DESIGN AND TOOL SOLUTIONS FOR ENERGY-EFFICIENT RELIABLE MONOLITHIC 3D ICS

A Dissertation
Presented to
The Academic Faculty

By

Kyungwook Chang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2019

Copyright © Kyungwook Chang 2019

DESIGN AND TOOL SOLUTIONS FOR ENERGY-EFFICIENT RELIABLE MONOLITHIC 3D ICS

Approved by:

Dr. Sung Kyu Lim, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Arijit Raychowdhury
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Saibal Mukhopadhyay
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Hyesoon Kim
College of Computing
Georgia Institute of Technology

Dr. Saurabh Sinha
Arm Research
Arm Inc.

Date Approved: March 18, 2019

To my parents Dr. Dooyung Chang and Dr. Euy Hyun Suh,

my brother Kyungjoon Chang,

my wife Hyekyoung Sung, and my daughter Claire Yunjin Chang

ACKNOWLEDGEMENTS

My journey to the doctoral degree is a series of marvelous experiences. Had it not been for the support and guidance from many individuals which have impacted the academic and social aspect of my life, I would not have completed the journey successfully. I would like to use this opportunity to show my appreciation to all those who have helped me along the way.

First of all, I would like to give my deepest gratitude to my advisor, Dr. Sung Kyu Lim, for his superb guidance and advice. He have helped me pursue the highest academic degree in one of the world's best institutions in this field. His insight, passion, and trust on me have enlightened me and greatly inspired my research.

I would like to express my heartfelt thank to Dr. Saurabh Sinha, my best mentor. He have helped my research with the continued support and invaluable guidance. His research experience and enthusiasm have significantly impacted my professional development.

I would like to thank Dr. Arijit Raychowdhury, Dr. Saibal Mukhopadhyay, and Dr. Hyesoon Kim for being the committee members and spending their valuable time for the invigorating discussions and feedbacks on my work.

My internship in Arm have helped me grow my insight on industrial researches. I thank Dr. Brian Cline and Dr. Greg Yeric for giving me the opportunity, and want to express my gratitude to Raney Southerland, Michael Doherty, Dr. Shidhartha Das, and Dr. Xiaoqing Xu for their aids and feedbacks on my research with Arm. I also appreciate Dr. Yu Cao and Dr. Jae-Sun Seo motivating and guiding me to new concepts on deep neural network, and want to thank to Dr. Dusan Petranovic for his valuable comments on my research.

I would like to show my sincere thanks to the past and current fellow graduate students in GTCAD Laboratory at Georgia Institute of Technology: Dr. Shreepad Panth, Dr. Taigon Song, Dr. Yarui Peng, Dr. Sandeep Samal, Bon Woong Ku, Kartik Acharya, Neela Lohith, Rakesh Perumal, Anthony Agnesina, Sai Pentapati, Jinwoo Kim, Da Eun Shim,

Jee Hyun Lee, Yi-Chen Lu, Lingjun Zhu, Gauthaman Murali, and Chengjia Shao for their great help and exciting discussions on research ideas. I would like to extend my thanks to Deepak Kadetotad for having our great collaborations.

I am thankful to my best friends in S. Korea, Bhumgey Lee, Kyudae Kim, Hanwool Leem, and Kyungjin Lee, who have supported me from my home country and encouraged me during my long journey. I also appreciate all my friends in the United States who have encouraged me, shared concerns, and made me feel like home away from home.

I would like to express my gratitude from the bottom of my heart to my parents, Dr. Dooyung Chang and Dr. Euy Hyun Suh, and my brother, Kyungjoon Chang. They have always been on my side, and their unconditional love, invaluable support, and indescribable devotion have guided and helped me become the person who I am. Without their support, I would not definitely have finished this long and challenging journey.

Lastly, I am particularly indebted to my wife, Hyekyoung Sung, and my young daughter, Claire Yunjin Chang, for their tremendous love, encouragement, and patience which cannot be mentioned in words. They have given me the strength and pleasure, which has been a constant source of motivation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The objective of this dissertation is to analyze and identify the benefits and challenges of energy-efficient and reliable monolithic 3D (M3D) ICs, and to develop physical design and tool solutions to address the challenges. The physical design and tool challenges of M3D ICs are addressed with categorizing them into four major themes, high-performance and low-power M3D ICs, new M3D IC design flow, power supply integrity of M3D ICs, and M3D ICs for deep neural network (DNN) hardware. In the first theme, the performance and power benefits of M3D ICs in advanced technology nodes are analyzed, and the optimization methodologies are presented to maximize the performance and to minimize the power consumption. In the second theme, an in-depth analysis on the power benefits of M3D ICs across technology nodes is performed, and a new M3D IC design flow is devised based on the observations. For the third theme, power supply integrity issues of M3D stacking technology are addressed, and the optimization methodologies are presented. Lastly, the challenges in DNN hardware and the impact of M3D ICs on the hardware are examined as implementing low-power and high-performance DNN hardware is known to be difficult albeit they are widespread and powerful in recognition tasks.

CHAPTER 1
INTRODUCTION

The objective of this research is to analyze and identify the benefits and challenges of energy-efficient and reliable monolithic 3D (M3D) ICs, and to develop physical design and tool solutions to address the challenges.

As technology scaling faces its physical limits in channel length scaling, degrading process variations, lithography constraints, increased parasitics, and rising manufacturing costs, M3D stacking technology takes center stage in continuing Moore's law. In M3D stacking technology, the devices are fabricated onto multiple tiers sequentially with nano-sized monolithic inter-tier vias (MIVs), which connect the topmost metal layer of the bottom tier and the bottommost metal layer of the top tier as shown in Figure 1.1. Because MIVs are extremely small, they can achieve much higher vertical integration density and lower RC parasitics compared to through-silicon vias (TSVs). Owing to the enhancement of fabrication technology, one can harness the true benefit of M3D ICs with fine-grained vertical integration [1].

M3D ICs show manifold advantages over conventional 2D ICs by utilizing short vertical connections instead of using long wires in the xy-plane, offering lower power consumption and higher performance. However, the 3D nature casts physical design and tool challenges.

1.1 Challenges in Monolithic 3D ICs

In this dissertation, the physical design and tool challenges of M3D ICs are addressed by categorizing them into four major projects.

In the first project, the performance and power benefits of M3D ICs in advanced technology nodes are analyzed, and the optimization methodologies are presented to maximize

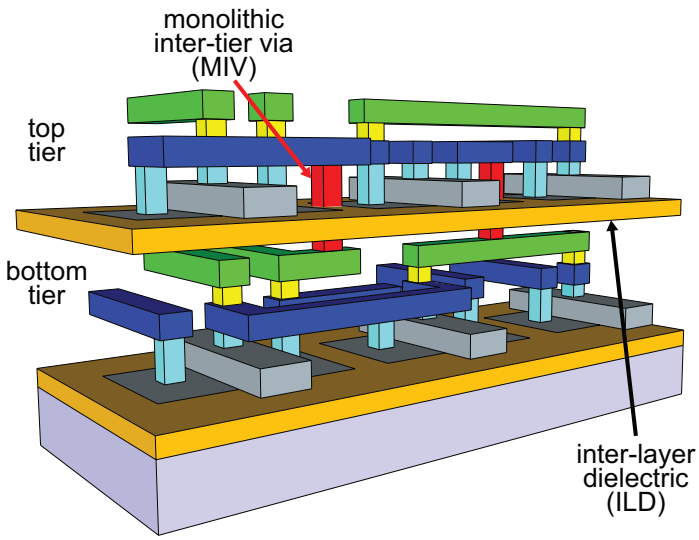

Figure 1.1: An example of M3D IC structure

the performance and to minimize the power consumption. Although various studies have been done to increase power savings of M3D ICs, efforts to improve their performance are rarely made. The performance of an IC is becoming more and more important in the modern world due to the high demand on computing and storage resources, which necessitates comprehensive studies on optimizing the performance of M3D ICs. In addition, devices have transitioned planar MOSFETs to 3D FinFETs in order to counteract the limits of degrading short channel effects, process variations, and reliability degradation. The device transition affects the power benefits of M3D stacking technology, but the impact has not yet been investigate thoroughly and accurately.

In the second project, an in-depth analysis on the power benefits of M3D ICs across technology nodes is performed, and a new M3D IC design flow is devised based on the observations. Because current commercial electronic design automation (EDA) tools do not support 3D placement of cells, several recent studies present methodologies to implement M3D ICs using commercial 2D tools. However, most of these studies implement each tier in a 3D IC separately after partitioning, which is prone to inaccurate buffer insertion especially in advanced technology nodes. Therefore, an M3D IC design flow, which implements and optimizes multiple tiers simultaneously in a single emulated 2D design, is

required.

For the third project, power supply integrity issues of M3D stacking technology are addressed, and the optimization methodologies are presented. Multiple device layers and reduced footprint make M3D ICs suffer from higher power density, which in turn, raise voltage drop on power rails of M3D ICs. In order to prevent functional failures and performance degradation due to worsened power supply integrity with lower supply voltage, faster operating clock frequency, and higher power density, power delivery networks (PDNs) of M3D ICs should be analyzed and optimized.

In the last project, the challenges in deep neural network (DNN) hardware and the impact of M3D ICs are examined as implementing low-power and high-performance DNN hardware is known to be difficult albeit they are widespread and powerful in recognition tasks. In order to resolve the high demand on computing and memory usage, and an extensive wire connections among neurons, studies, which encompass the influence of key parameters in DNN hardware including DNN architecture choices and underlying workloads as well as the physical design optimization, are required.

1.2 Contributions

The main contributions of this dissertation are as follows:

High-Performance and Low-Power Monolithic 3D ICs

- The performance benefits of M3D ICs over 2D ICs are quantified
- In-depth analysis on the key factors which affect the performance is performed
- Methodologies to maximize the performance benefits of M3D ICs are presented
- A predictive $7nm$ process design kit (PDK) based on FinFETs is developed, and its standard cell libraries with high performance (HP) and low standby power (LSTP) cells are created

- The impact of M3D stacking technology on the power consumption at $7nm$ FinFET technology node is examined
- Guidelines and a new partitioning methodology to maximize the power benefit of M3D ICs in advanced technology nodes are presented

New Monolithic 3D IC Design Flow

- M3D ICs across multiple technology nodes, namely, foundry $28nm$, 14/16nm and a predictive $7nm$ technology nodes, are implemented and analyzed with a commercial microprocessor as the benchmark
- An extensive set of results explaining the factors that impact power savings in M3D stacking technology is presented and analyzed
- A novel M3D IC design flow that incorporates design and micro-architecture insight to guide the partitioning scheme is presented
- The new M3D IC design flow is partition-scheme agnostic, hence, making it an ideal platform to evaluate different partitioning schemes
- The new M3D IC design flow effectively reduces standard cell area as well as wire-length compared to 2D ICs, resulting in significantly better power saving compared to existing M3D IC design flows

Power Supply Integrity of Monolithic 3D ICs

- The benefits and challenges of PDNs in M3D ICs are investigated
- A system-level PDN circuit of M3D ICs as well as 2D ICs is modeled, and an in-depth analysis on both static and dynamic behavior of the PDN is performed
- A resonance frequency analysis and an in-rush current study for M3D ICs, which show the frequency- and time-domain response of the PDNs, are presented

- Two optimization methodologies for the PDNs of M3D ICs, top-tier cell repositioning and asymmetric top- and bottom-tier M3D PDN, are presented, based on the lessons learned from the analysis

Monolithic 3D ICs for Deep Neural Network Hardware

- The impact of M3D ICs on DNN architectures with different granularity in sparsity is investigated
- The impact of tier partitioning in M3D ICs to better handle memory blocks is studied
- Feed-forward classification and pseudo-training workloads are examined thoroughly to investigate their impact on power reduction
- An in-depth analysis on the performance benefit of DNN M3D ICs over their 2D counterparts is demonstrated
- Key guidelines on optimal architectural and physical design decisions for DNN hardware are presented

1.3 Organization

This dissertation is orgainized as follows:

- In Chapter 2, the performance and power benefits of M3D ICs are analyzed, and their optimization methodologies are presented
- In Chapter 3, The power benefit trends of M3D stacking technologies across multiple technology nodes are presented, and a new M3D IC design flow is presented, which maximizes the power benefits in advanced technology nodes
- In Chapter 4, power supply integrity of M3D ICs are examined with an in-depth analysis, which encompasses static and dynamic rail analysis, and time- and frequency-domain response, and the PDN optimization methodologies for M3D ICs are proposed

- In Chapter 5, The influence of key parameters in DNN hardware architecture and M3D stacking technology are investigated, and the optimization methodologies to improve performance and power savings are presented
- In Chapter 6, the conclusions of this dissertation are presented

CHAPTER 2

HIGH-PERFORMANCE AND LOW-POWER MONOLITHIC 3D ICS

2.1 Motivation and Background

2.1.1 Monolithic 3D IC Performance Improvement

Previous studies have explored and shown the performance optimization for 3D ICs [2, 3]. However, these studies have focused primarily on investigating TSV-based 3D ICs, which inevitably show much lower vertical integration densities due to the large micron-scale size and the keep-out zone (KOZ) of TSVs. TSV-based 3D ICs fail to fully benefit from 3D IC stacking technology because of the lower vertical integration density.

There has been active research on power benefits of M3D ICs over 2D counterparts, offering quantifiable power savings [4, 5]. However, none of these works attempted performance optimization for M3D ICs, but instead focused only on power benefits. The authors of [6] proposed methodologies to improve the performance of M3D ICs by vertically stacking diffusion area of cells, but their methods involve custom cell designs, which require excessive effort and time.

With extensive data and memory requirements, high performance has become a game changing design factor in the modern world. Therefore, comprehensive studies on optimizing the performance of M3D ICs are required.

2.1.2 Power Saving of Monolithic 3D ICs in Advanced Technology Nodes

In advanced technology nodes, devices have transitioned from planar MOSFETs to 3D FinFETs in order to counteract the limits of degrading short channel effects, process variations, and reliability degradation.

While M3D stacking technology based on planar MOSFETs has been studied actively,

FinFET implementations have not been widely explored. The authors of [5] demonstrated a computer-aided design (CAD) methodology for gate-level M3D ICs, but it is based on $28nm$ process which does not utilize FinFETs. Recently, a study on the benefits of transistor-level M3D ICs on a $7nm$ cell library has been investigated in [7]. However, their work used the $7nm$ FinFET model only for generating timing and power metrics for a few low drive-strength cells. They extrapolated results from the metrics, not considering the structure and effects of FinFETs during cell design. This simple extrapolation is prone to inaccuracies, and it is important to consider FinFET-based cell design and model scaled wire RC parasitics to obtain a complete and accurate estimate of the impact of M3D ICs utilizing FinFETs.

In order to examine the benefit and trade-offs involved in using M3D ICs at the end of silicon scaling, a predictive PDK and its cell libraries are needed, and the power consumption of M3D ICs should be analyzed and optimized based on the PDK and cell libraries.

2.2 Tool Solutions for High-Performance Monolithic 3D ICs

In this work, comprehensive studies on the factors that impact the performance benefit of M3D ICs are performed, and methodologies to further improve the performance are presented. The findings in this work are supported with an in-depth analysis, and the presented methodologies help raise maximum achievable clock frequencies of M3D ICs

2.2.1 Existing Full-Chip Monolithic 3D IC Design Flow

Design Methodology

Due to the lack of ability of current commercial EDA tools to place cells in 3D space, **shrunk-2D design flow** [5] uses dimensional shrinking techniques with 2D commercial EDA tools to implement two-tier M3D ICs. Figure 2.1 summarizes the design flow.

The first step involves floorplanning hard macros like memory blocks in M3D ICs. They can be placed in either or both the tiers. Next, the floorplan for the corresponding *shrunk-*

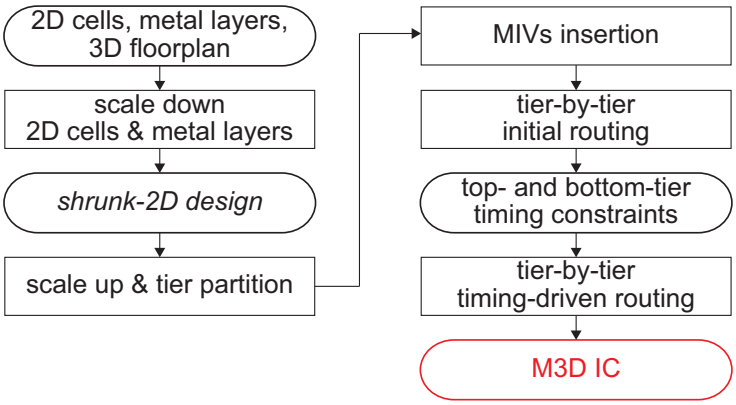

Figure 2.1: An existing CAD methodology to implement M3D ICs [5]

2D design is created, which includes creating placement blockages on locations where both tiers have hard macros and partial placement blockages if only one of the tiers contains hard macros on the locations. In the final M3D IC, the cells in the partial placement blockages will be assigned to the tier which does not contain hard macros.

An M3D IC is implemented in a footprint half the size of its 2D counterpart by utilizing two tiers. In order to make use of 2D commercial EDA tools to place all cells and wires on the halved footprint, the width and height of standard cells are scaled down by $1/\sqrt{2}$. The width and pitch of metal wires are scaled down as well to compensate for the shrunk footprint.

With the shrunk cells and wires and the floorplan of a *shrunk-2D design*, all the design stages including placement, post-placement optimization, clock-tree-synthesis (CTS), post-CTS optimization, routing, and post-route optimization are performed using Cadence® Innovus™, creating a shrunk version of the 2D design, *shrunk-2D design*.

The cells in the *shrunk-2D design* are, then, scaled up to the original size creating overlaps among the cells in the design. In order to remove the overlap, the design is divided into multiple square bins on the xy-plane, and for each bin, the cells are partitioned into two tiers using an area-balanced min-cut partitioning algorithm, so that half of the cells in the bin are assigned on the bottom tier, and the other half on the top tier, determining the z-location of each cell. After partitioning the cells for all bins, the cells in each tiers

are legalized in order to remove overlap remained even after partitioning, deriving the final xy-location of the cells.

The back-end-of-line (BEOL) metal stack is duplicated to account for the metal layers in each tier. Additionally, the cells are annotated with their respective tiers. Every cell has its pins on the metal layers of their corresponding tier. With the duplicated metal stacks, xy-location, and annotated pins of the cells, the design is routed using Cadence® Innovus™. The location of MIVs are determined by the locations of the vias between the topmost metal layer of the bottom tier and the bottommost metal layer of the top tier.

Then, using the netlists of each tier and MIV location, trial-route is performed on each tier, and the initial top- and bottom-tier designs are fed to Synopsys PrimeTime® to obtain timing constraints for each tier. Once the timing constraints are determined, timing-driven routing is performed for each tier design with the timing constraints, resulting in the final top- and bottom-tier designs of the M3D IC.

Limitations

Inaccurate Parasitic Estimation In the baseline **shrunk-2D design flow**, cell drive-strength, cell xy-location, and buffer insertion are determined while implementing *shrunk-2D designs*, and they are not changed afterwards. Since cell placement and optimizations are based on estimated wire RC parasitics, it is crucial for *shrunk-2D designs* to estimate wire RC parasitics of the corresponding M3D ICs correctly as described in Equation (2.1).

$$\begin{aligned} R_{S2D} = R_{M3D}, \quad C_{S2D,GND} = C_{M3D,GND} \\ C_{S2D,X,h} = C_{M3D,X,h}, \quad C_{S2D,X,v} = C_{M3D,X,v}, \end{aligned} \tag{2.1}$$

where R and C_{GND} represent the resistance and ground capacitance of a wire. $C_{X,h}$ and $C_{X,v}$ are the horizontal (i.e., the xy-plane) and vertical (i.e., the z-axis) coupling capacitance, respectively. $S2D$ refers to *shrunk-2D designs* while $M3D$ refers to the corresponding M3D ICs.

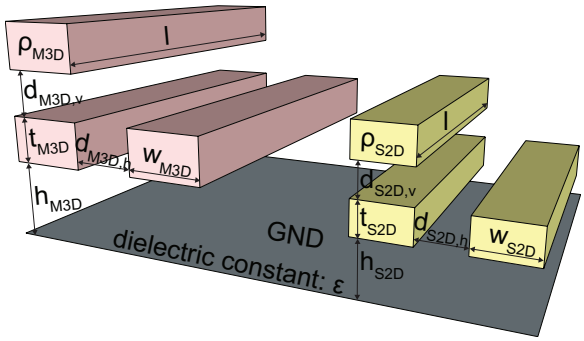

Figure 2.2: Comparison of wire geometries in an M3D IC (pink) and its *shrunk-2D design* (yellow)

Current commercial EDA tools utilize RC parasitic look-up tables (i.e., qrcTechFile in Cadence® tools or nxtgrd in Synopsys® tools) to determine the resistance and capacitance of wires based on geometric and material parameters of wires. However, the dimensional shrinking of wires makes *shrunk-2D designs* suffer from inaccurate RC parasitic estimation of the final M3D ICs.

The resistance, and the ground and coupling capacitances of the wires are described by the following equations.

$$\begin{aligned} R = \rho \cdot l/(t \cdot w), \quad C_{GND} = \epsilon \cdot (w \cdot l)/h \\ C_{X,h} = \epsilon \cdot (t \cdot l)/d_h, \quad C_{X,v} = \epsilon \cdot (w \cdot l)/d_v, \end{aligned} \tag{2.2}$$

where ρ and ϵ indicate the resistivity of wires and the dielectric constant of dielectric layers. w, h, and t represent the width, height from the ground, and thickness of the wires, whereas d_h and d_v are the horizontal and vertical spacing of the wires, respectively.

Figure 2.2 compares wires in an M3D IC, and the corresponding wires in the *shrunk-2D design* with the same length, l. The following constraints are applied on the width and horizontal spacing of the wires in the *shrunk-2D design* due to the dimensional shrinking on the xy-plane.

$$w_{S2D} = w_{M3D}/\sqrt{2}, \qquad d_{S2D,h} = d_{M3D,h}/\sqrt{2} \tag{2.3}$$

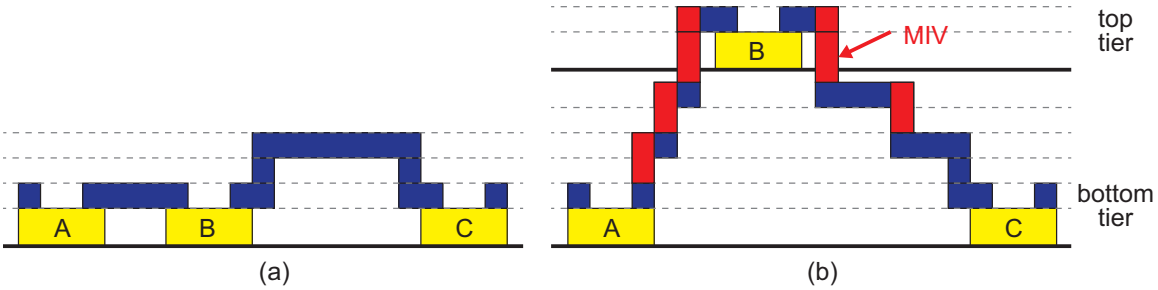

Figure 2.3: Inter-tier routing overhead (red) resulting from assigning cell B on the top tier while transforming (a) a *shrunk-2D design* to (b) the M3D IC.

From Equation (2.2) and Equation (2.3), Equation (2.4) is derived, which describes the RC parasitics estimated in the baseline **shrunk-2D design flow** as it assumes all parameters (except the width and horizontal spacing) are the same in a *shrunk-2D design* and the corresponding M3D IC.

$$\begin{aligned} R_{S2D-base} &= \sqrt{2} \cdot R_{M3D}, \quad C_{S2D-base,GND} = C_{M3D,GND}/\sqrt{2}, \\ C_{S2D-base,X,h} &= \sqrt{2} \cdot C_{M3D,X,h}, \quad C_{S2D-base,X,v} = C_{M3D,X,v}/\sqrt{2} \end{aligned} \tag{2.4}$$

The above equation indicates that *shrunk-2D designs* implemented with the baseline **shrunk-2D design flow** estimate the RC parasitic of their M3D ICs incorrectly.

Inter-Tier Routing Overhead As cells are not assigned to tiers in *shrunk-2D designs*, the timing closure in a *shrunk-2D design* does not take into account the routing resources necessary to connect cells in different tiers through MIVs (i.e., inter-tier routing overhead). The inter-tier routing overhead in M3D ICs makes the delay of the timing paths crossing the tiers longer than the *shrunk-2D designs*. As the area-balanced min-cut partitioning algorithm in the baseline **shrunk-2D design flow** is timing-agnostic, dividing cells in critical timing paths into two tiers can degrade the performance of the M3D ICs as shown in Figure 2.3.

The inter-tier routing overhead induced by MIVs also negatively affects the quality of the clock tree of M3D ICs. Figure 2.4 illustrates the degraded clock skew in an M3D IC due to MIV insertion. The clock signal at the start point (cell C) and the end point (cell

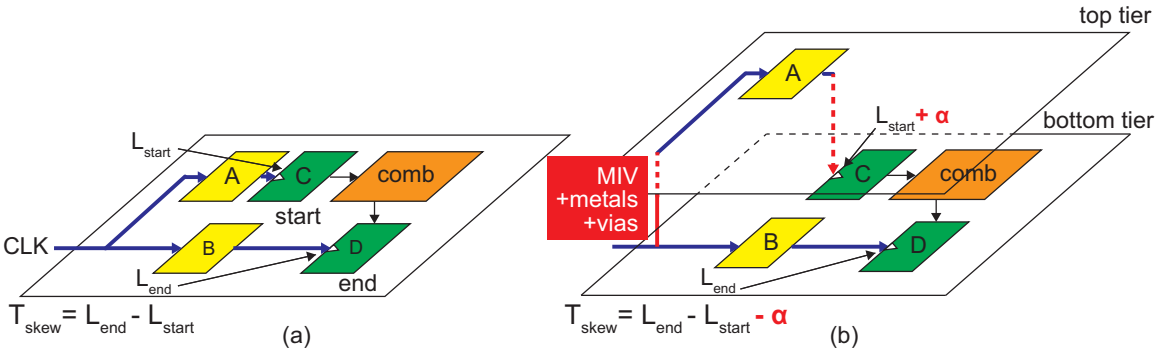

Figure 2.4: Impact of inter-tier routing overheads on clock skew. (a) The clock tree in a *shrunk-2D design*, and (b) the degraded clock skew due to the inter-tier routing overhead in the M3D IC (= MIV + metals + vias).

D) of the timing path are fed from clock buffers, cell A and cell B, respectively. In the *shrunk-2D design*, commercial EDA tools minimize T_{skew} in Figure 2.4 (a), maximizing the timing budget of the timing paths. However, if cell A is placed on the top tier, inter-tier routing overheads are introduced to the clock path to cell C. This, in turn, increases the clock latency at the clock pin of the start point (cell C) by α (as shown in Figure 2.4 (b)), reducing the timing budget of the timing path.

2.2.2 Computer-Aided Design Solutions to Improve Performance of Monolithic 3D ICs

In order to quantify the performance of a design, the slack which describes the timing closure of a timing path in a design, and worst negative slack (WNS) which determines the maximum performance of the design are employed. A negative WNS indicates that the design fails to close timing, and a slower clock frequency is required. The slack of a timing path is described with Equation (2.5).

$$\text{slack} = T_{CP} + T_{skew} - T_{setup} - D_{path}, \tag{2.5}$$

where T_{CP} and T_{skew} indicate the clock period and clock skew between the start and end point of a timing path, respectively. T_{setup} is the setup time of the end point, and D_{path} represents the total path delay of the combinational logic in the timing path.

Performance optimization for M3D ICs is performed with three techniques: parasitic

adjustment during implementing *shrunk-2D designs*, path-based tier partitioning and clock buffer tier partitioning while transforming *shrunk-2D designs* into the M3D ICs.

Parasitic Adjustment

In order for *shrunk-2D designs* to accurately estimate the RC parasitics of the wires in the M3D ICs while using the dimensional scaling technique, the RC parasitic look-up tables should be updated by modifying the geometric and material parameters (i.e., the terms in the right-hand side in Equation (2.2)), so that the look-up tables reflect the shrunk wires to determine wire RC parasitics. One of the solutions for Equation (2.1) with the constraints in Equation (2.3) is described in the following equations.

$$\begin{aligned} &\rho_{S2D} = \rho_{M3D}/2, \qquad \epsilon_{S2D} = \epsilon_{M3D}, \qquad h_{S2D} = h_{M3D}/\sqrt{2}, \\ &t_{S2D} = t_{M3D}/\sqrt{2}, \quad d_{S2D,v} = d_{M3D,v}/\sqrt{2} \end{aligned} \tag{2.6}$$

In this parasitic adjustment (PA) methodology, Equations (2.6) is used to generate an RC parasitic look-up table, which is fed to commercial EDA tools to calculate the RC parasitics of the wires in a design. With the adjusted RC parasitic look-up table, a *shrunk-2D design* is implemented. It is important to note that the adjusted RC parasitic look-up table is used only for *shrunk-2D designs*, and the original RC parasitic look-up table is used to calculate the wire RC parasitics of the final M3D ICs.

Table 2.1 presents the error rate of the RC parasitics of a *shrunk-2D design* from its M3D ICs, comparing the baseline **shrunk-2D design flow** and the one with parasitic adjustment. The table clearly shows that the parasitic adjustment decreases the error rate of the RC parasitics of the *shrunk-2D design* to 11.6% and 4.3% for the total capacitance and resistance of the design, respectively. Although the parasitic adjustment technique reduces the error rate of the RC parasitics compared to the baseline **shrunk-2D design flow**, errors still remain because the cells in a *shrunk-2D design* are partitioned into the top and bottom tiers, and the wires are re-routed with the partitioned design.

With the optimized *shrunk-2D design* implemented with the parasitic adjustment tech-

Table 2.1: Parasitic error rate of a *shrunk-2D design* from its M3D IC, comparing the baseline **shrunk-2D design flow** and the one with parasitic adjustment in an LDPC design

parameters		*shrunk-2D design*	M3D IC	ERR%
baseline **shrunk-2D**	pin cap (pF)	137.6	137.6	0.0%
	wire cap (pF)	250.6	202.3	23.9%
	total cap (pF)	**388.2**	**339.9**	**14.2%**
	resistance ($M\Omega$)	**8.663**	**7.470**	**16.0%**
shrunk-2D w/ parasitic adjustment	pin cap (pF)	118.8	118.8	0.0%
	wire cap (pF)	227.5	191.4	18.9%
	total cap (pF)	**346.3**	**310.2**	**11.6%**
	resistance ($M\Omega$)	**5.039**	**4.833**	**4.3%**

nique, the next two techniques describe performance optimization methodologies while transforming *shrunk-2D designs* into the M3D ICs.

Path-based Tier Partitioning

A path-based tier partitioning algorithm [8] is employed to prevent from increasing the delay due to MIV insertion on the critical paths. This algorithm first pre-places all cells in a number of critical timing paths on a single tier to avoid inter-tier routing overheads, and then performs the area-balanced min-cut partitioning with the remaining cells. The algorithm was originally devised to overcome inter-tier degradation of M3D ICs, but it is used to improve the performance of M3D ICs in this work by minimizing inter-tier routing overheads.

Clock Buffer Tier Partitioning

While the aforementioned two techniques focus on reducing D_{path} in Equation (2.5), the slack of a timing path can also be improved by preventing negative T_{skew}. The following describes the clock skew of a timing path.

$$T_{skew} = L_{end} - L_{start}, \tag{2.7}$$

where L_{end} and L_{start} are the clock latency at the clock pin of the end and start point of a timing path. Increasing L_{end} may improve slack for a timing path, but it imposes a tighter

timing constraint on the subsequent path, which uses the end point of the current timing path as its start point. Therefore, it is important to obtain almost uniform clock latencies, making T_{skew} nearly zero for all timing paths if useful skews are not considered.

To prevent clock skew degradation, the methodology proposed in [9] is utilized, which was originally devised to minimize the clock power consumption of M3D ICs. The methodology partitions clock cells onto a single tier, which helps minimize MIV utilization in clock paths. However, placing all clock cells including flip-flops may cause significant area skew between tiers, increasing wire congestion in the tier with a higher cell area. The wire congestion increases the wire-length of M3D ICs, resulting in performance degradation. Therefore, only clock buffers are assigned to a single tier, and let flip-flops be free to be partitioned in either tier, so that every clock path in a clock tree utilizes at most one MIV, minimizing inter-tier routing overheads.

Experimental Setup

NanGate FreePDK45 Open Cell Library is used for implementing 2D and M3D ICs. Rocket-core and ride-core, which are based on the RISC-V instruction set architecture, and advanced encryption standard 128 bit (AES-128) and low-density parity-check (LDPC) from OpenCores are used as test vehicles. The 2D and M3D ICs of the four benchmarks are implemented sweeping the target frequencies in order to obtain their maximum frequencies. The footprint and floorplan of each design are customized to maximize their performance without design rule check (DRC) violations, and kept constant during frequency sweeps.

Monolithic 3D IC Performance Improvement Results

Table 2.2 compares the 2D and M3D ICs with their maximum achievable clock frequencies. As M3D ICs have cells on two tiers, the footprint of the M3D ICs are $\sim$50% smaller than their 2D counterparts. As a result, the wire-length reduction of the critical timing path reaches 53.4% due to shorter distances between cells in the M3D ICs. The reduced wire-

Table 2.2: Maximum performance comparison of 2D and M3D ICs with the performance improvement techniques. The percentage values in the M3D ICs are with respect to their 2D counterparts. Area values are in in μm^2, frequency in MHz, power in mW, energy-delay-product (EDP) in $pJ \cdot ns$, length in μm, capacitance in pF, resistance in Ω, and time in ns.

parameters	rocket-core			ride-core			AES-128			LDPC		
	2D	M3D		2D	M3D		2D	M3D		2D	M3D	
full-chip power, performance, and area (PPA)												
footprint	0.384	0.188	(-51.0%)	0.523	0.256	(-52.4%)	0.390	0.191	(-51.0%)	0.228	0.112	(-51.0%)
eff freq	**783.3**	**905**	**(15.5%)**	**524**	**550.0**	**(5.0%)**	**1,388**	**1,585**	**(14.2%)**	**716**	**828**	**(15.6%)**
total power	156.7	175.2	(11.8%)	147.8	149.8	(1.4%)	177.0	207.9	(17.5%)	219.4	252.1	(14.9%)
EDP	**255.4**	**214.0**	**(-16.2%)**	**539.3**	**495.9**	**(-8.1%)**	**91.9**	**82.8**	**(-10.0%)**	**428.0**	**368.2**	**(-14.0%)**
critical timing path												
target freq	813	938	(15.4%)	530	562	(5.9%)	1,375	1,750	(27.3%)	750	875	(16.7%)
wire-length	900.4	419.3	(-53.4%)	865.1	816.9	(-5.6%)	91.9	95.5	(3.9%)	1,519.9	987.3	(-35.0%)
std. cell area	53.7	34.8	(-35.1%)	76.1	72.6	(-4.5%)	18.9	17.4	(-7.8%)	46.3	40.5	(-12.6%)
wire cap	0.110	0.063	(-42.9%)	0.167	0.138	(-16.9%)	0.021	0.022	(2.3%)	0.236	0.191	(-19.2%)
pin cap	0.250	0.098	(-60.7%)	0.263	0.283	(7.5%)	0.031	0.041	(32.7%)	0.169	0.125	(-26.1%)
resistance	4,234	3,242	(-23.4%)	5,287	5,438	(2.9%)	1,109	1,090	(-1.7%)	5,581	3,647	(-34.7%)
setup time	0.016	0.031	(92.6%)	0.015	0.019	(30.4%)	0.037	0.037	(0.5%)	0.000	0.028	(-)
clk skew	-0.011	-0.131	(1,104.6%)	0.071	-0.068	(-195.9%)	-0.039	-0.045	(14.8%)	-0.214	-0.068	(-68.1%)
delay	1.249	0.943	(-24.6%)	1.966	1.732	(-11.9%)	0.644	0.548	(-14.9%)	1.183	1.113	(-5.9%)
WNS	-0.046	-0.039	(-15.7%)	-0.024	-0.038	(58.3%)	0.007	-0.059	(-1,004.8%)	-0.063	-0.066	(3.6%)

length helps reduce the wire-load of the timing paths, which in turn, decreases standard cell area, showing up to 35.1% reduction. The reduced wire-length decreases the wire capacitance and resistance of the critical timing path, while the reduced pin capacitance is attributed to the standard cell area savings of the timing path. The reduced wire RC parasitics help lower the delay of the nets in the critical timing path. Since the wire-loads at the output of the cells are reduced, the cell delays are also decreased, offering up to a 24.6% total delay reduction. As the delay of the timing path is decreased, the effective frequency, the highest clock frequency at which the design can operate, is increased up to 15.6%.

It is important to note that the slack of a timing path is determined not only by D_{path} but also by T_{CP}, which is smaller in the M3D ICs. Therefore, a worse WNS in M3D ICs does not necessarily mean that the timing closure is worse than its 2D counterpart.

Table 2.2 also shows that the performance improvement of ride-core is lower than the other benchmarks, showing only 5.0% improvement. This is mainly due to the complexity of the timing paths in the design. Figure 2.5 shows the timing path wire-length distribution of the four benchmarks. Ride-core has a higher number of timing paths with large wire-length, which makes it difficult for EDA tools to close timing. In those designs, cells tend to be clustered to each other, providing less room for reducing wire-length and standard cell area of the critical timing path in the M3D ICs.

Figure 2.6 shows the benefits of M3D ICs on the timing closure, comparing the slack distribution of the timing paths in the 2D and M3D ICs. The 2D IC suffers from a large number of the timing paths with near-zero and negative slack, whereas the timing paths of the M3D IC show a larger number of higher positive slack. The figure indicates that the reduced wire-length and standard cell area of the timing paths of the M3D ICs help ease timing closure, hence, offering performance improvement.

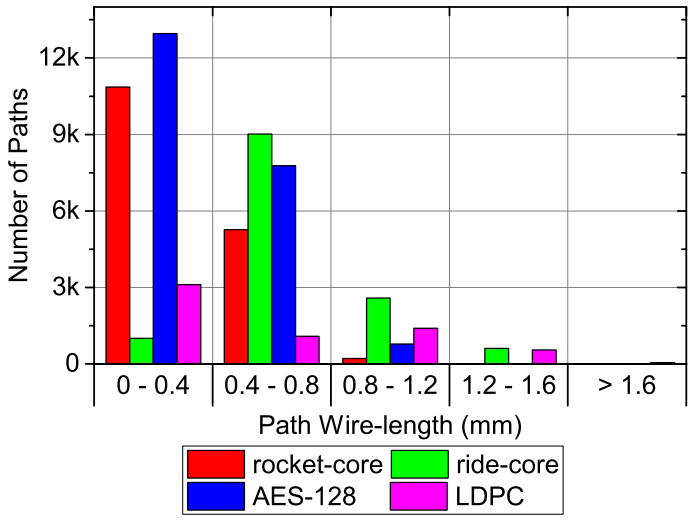

Figure 2.5: Wire-length distribution of the timing paths in 2D and M3D ICs

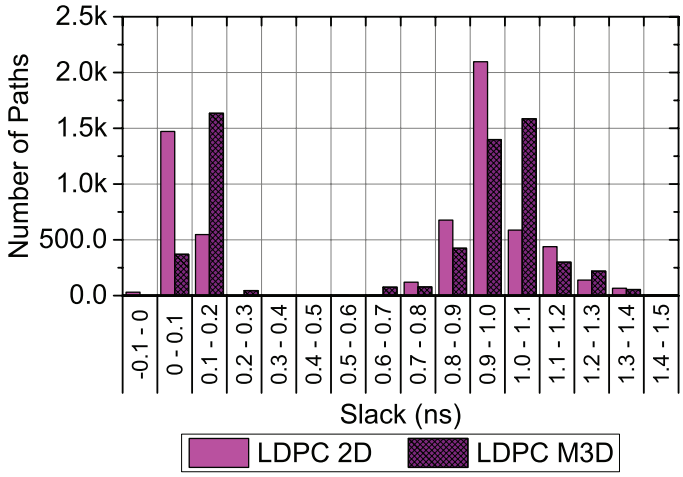

Figure 2.6: Slack distribution comparison between 2D and M3D ICs in an LDPC design

On Performance Improvement Methods

The impact of each performance improvement methodologies presented is shown in Table 2.3. In all benchmarks, the M3D ICs implemented with parasitic adjustment outperform the ones using the baseline **shrunk-2D design flow**, showing up to 8.0% performance improvement. The baseline **shrunk-2D design flow** fails to utilize enough number of buffers and drive-strength cells on the critical timing paths due to inaccurate RC parasitic estimation, reducing the performance improvement. The benefit of the adjusted parasitics is

Table 2.3: Impact of the parasitic adjustment, path-based tier partitioning, and clock buffer tier partitioning on the performance of M3D ICs. Values are the performance improvement with respect to their 2D ICs.

techniques	rocket-core	ride-core	AES-128	LDPC
baseline **shrunk-2D**	10.2%	0.5%	11.5%	7.6%
+ parasitic adjustment	12.6%	3.6%	13.5%	**15.6%**
+ path-based tier partitioning	14.9%	**5.0%**	**14.2%**	10.7%
+ clock buffer tier partitioning	**15.5%**	4.4%	11.9%	10.1%

greater in wire-dominated circuits (e.g., LDPC design) as wire RC parasitic estimation significantly affects the cell placement.

The parasitic adjustment technique also helps reduce the power consumption of M3D ICs. Table 2.4 shows the iso-performance power metric comparison of the M3D ICs implemented using the baseline **shrunk-2D design flow** and the one with parasitic adjustment. Because the RC parasitics of the wires in *shrunk-2D designs* more accurately depict the final M3D ICs with parasitic adjustment, lower number of buffers and drive-strength cells are utilized which are in non-critical paths, resulting in lower power consumption.

In the path-based tier partitioning method, pre-placing cells in a number of critical timing paths improves the performance of the M3D ICs by reducing the inter-tier routing overheads along their critical timing paths. However, in benchmarks with a high number of timing paths with near-zero slack like LDPC design (as shown in Figure 2.6), pre-placing cells does not improve the performance. In those benchmarks, the partitioning method improves the timing closure of the selected timing paths. However, the method makes the area-balanced min-cut partitioning algorithm more difficult to minimize the number of MIVs. This increases the delay of the non-selected timing paths with near-zero slack, making them new critical timing paths.

The clock buffer tier partitioning technique offers a performance benefit to the rocket-core M3D IC. The sequential cells occupy 19.1% of the total cells in the 2D rocket-core, which is a higher percentage compared to the other benchmarks (ride-core = 9.3%, AES-128 = 8.1%, and LDPC = 3.0%). As the clock tree is more complex in sequential cell-

Table 2.4: Iso-performance power metric comparison of the M3D ICs implemented with the baseline **shrunk-2D design flow** and the one with parasitic adjustment. The percentage values in the M3D ICs are with respect to their 2D counterparts. Power values are in mW.

design	parameters	M3D IC			
		baseline **shrunk-2D**		**shrunk-2D** w/ parasitic adjustment	
rocket-core	net switching power	41.3	(-9.1%)	40.1	(-11.8%)
	internal power	112.4	(-1.4%)	110.4	(-3.1%)
	leakage power	3.0	(-2.7%)	2.9	(-6.8%)
	total power	**156.7**	**(-3.5%)**	**153.4**	**(-5.6%)**
ride-core	net switching power	53.6	(-16.7%)	49.6	(-23.0%)
	internal power	78.1	(-3.4%)	72.8	(-10.0%)
	leakage power	4.1	(-4.8%)	3.7	(-14.2%)
	total power	**135.9**	**(-9.2%)**	**126.1**	**(-15.7%)**
AES-128	net switching power	64.5	(-7.6%)	61.2	(-12.2%)
	internal power	100.2	(-1.7%)	95.3	(-6.6%)
	leakage power	3.6	(-3.0%)	3.2	(-13.5%)
	total power	**168.3**	**(-4.1%)**	**159.7**	**(-9.0%)**
LDPC	net switching power	79.0	(-41.9%)	71.4	(-47.5%)
	internal power	71.6	(-21.4%)	63.5	(-30.2%)
	leakage power	2.1	(-22.2%)	1.8	(-32.7%)
	total power	**152.6**	**(-33.5%)**	**136.7**	**(-40.5%)**

dominated designs, their M3D ICs benefit more from the methodology, decreasing the inter-tier routing overheads. On the other hand, it negatively affects the performance of combinational cell-dominated circuits since it is prone to increase the number of MIVs, worsening the delay of the timing paths.

2.3 Monolithic 3D IC Power Optimization in Advanced Technology Nodes

In this work, the power benefits of M3D ICs in a $7nm$ FinFET technology node are investigated, and design guidelines in advanced technology nodes are presented. A predictive $7nm$ FinFET PDK and standard cell libraries using both HP and LSTP device technologies are developed based on NanGate FreePDK45 Open Cell Library using accurate dimensional, material, and electrical parameters from publications and a commercial-grade tool flow. Then, using the PDK and standard cell libraries, M3D ICs are implemented to gauge the impact of M3D stacking technology on power metrics. Design guidelines as well as a

new partitioning methodology to improve the quality of M3D ICs are presented.

2.3.1 $7nm$ Process Design Kit Generation

In order to properly evaluate the power benefits of M3D ICs on a $7nm$ FinFET technology node, the corresponding PDK is needed for standard cell design. Since an open-source $7nm$ FinFET PDK is not readily available to the research community, a predictive $7nm$ FinFET PDK and standard cell libraries are created and validated. Starting from NanGate FreePDK45, all technology parameters are scaled corresponding to a $7nm$ technology node. This section presents the procedure used to develop the predictive $7nm$ FinFET PDK.

Process Modeling and Process Design Kit Generation

A PDK is defined based on the minimum dimensions of each layer in the process and accurate modeling of the transistor and interconnect behavior.

Dimensional Scaling Table 2.5 shows the minimum dimensions and material properties assumed in the $7nm$ FinFET PDK. Channel length scaling has been less aggressive in sub-$45nm$ technology nodes and is no longer the primary parameter defining a technology node. However, contacted poly-pitch (CPP) and M1 pitch scale by about $0.7\times$ every node and are better indicators of expected area scaling. Based on industry trends and [10], the values of $35nm$ for M1 pitch and $48nm$ CPP are used for the $7nm$ FinFET PDK. To scale the $45nm$ layouts to $7nm$ dimensions, the geometric mean of the M1 pitch and CPP are utilized to obtain the scaling factor of 0.25[1].

For interconnect dimensions, all xy-dimensions of wires are scaled from NanGate FreePDK45 by the same scaling factor of 0.25, but the aspect ratios (i.e., thickness/width) are set to 2 based on ITRS projections. The dielectric thicknesses are scaled proportionately from NanGate FreePDK45.

[1]Due to precision problems with the EDA tools, the scaling factor is rounded to 2 decimal places.

Table 2.5: Key parameters in NanGate FreePDK45 and the predictive $7nm$ FinFET PDK

parameters		NanGate FreePDK45	predictive $7nm$ FinFET PDK	
V_{DD} (V)		1.1	0.7	(-36.4%)
L_G (μm)		0.0500	0.0125	(-75.0%)
M1 pitch (μm)		0.1400	0.0350	(-75.0%)
contacted poly pitch (CPP) (μm)		0.1900	0.0480	(-74.7%)
cell height (M1 track)		10TR	10TR	(-75.0%)
M1	width (μm)	0.0700	0.0174	(-75.1%)
	thickness (μm)	0.1300	0.0348	(-73.2%)
	dielectric thickness (μm)	0.2500	0.0673	(-73.1%)
	sheet resistance $(\Omega/\square)$	0.3800	1.8200	(378.9%)
VIA1	via resistance (Ω)	5.0000	36.4000	(628.0%)
M4	width (μm)	0.1400	0.0350	(-75.0%)
	thickness (μm)	0.2800	0.0700	(-75.0%)
	dielectric thickness (μm)	0.5700	0.1425	(-75.0%)
	sheet resistance $(\Omega/\square)$	0.2100	0.9070	(331.9%)
VIA4	via resistance (Ω)	3.0000	8.7200	(190.7%)
M7	width (μm)	0.4000	0.1000	(-75.0%)
	thickness (μm)	0.8000	0.2000	(-75.0%)
	dielectric thickness (μm)	1.6200	0.4050	(-75.0%)
	sheet resistance $(\Omega/\square)$	0.0750	0.0950	(26.7%)
VIA7	via resistance (Ω)	1.0000	0.8330	(-16.7%)
M9	width (μm)	0.8000	0.2000	(-75.0%)
	thickness (μm)	2.0000	0.4000	(-80.0%)
	dielectric thickness (μm)	4.0000	0.8000	(-80.0%)
	sheet resistance $(\Omega/\square)$	0.0300	0.0475	(58.3%)
VIA9	via resistance (Ω)	0.5000	0.2960	(-40.8%)

Interconnect Modeling A PDK requires accurate modeling of interconnect parameters such as conductor sheet resistance, via and contact resistance. Copper (Cu) is assumed to be used for metal layers, and the resistivity of M1 through M6 layers is determined to be $6.35\mu\Omega \cdot cm$, and $1.9\mu\Omega \cdot cm$ for M7 through M10, based on ITRS projections. One of the main reasons for the increased resistivity is increased scattering experienced at grain boundaries within the Cu wires [11]. Due to the increased resistivity and diminished cross-sectional area, the sheet resistances of the $7nm$ FinFET PDK are larger than those of NanGate FreePDK45 as shown in Table 2.5.

For vias, Cu is assumed for via material with Tantalum Nitride (TaN) barrier. A barrier is necessary between a Cu via and the corresponding dielectric layer in order to prevent

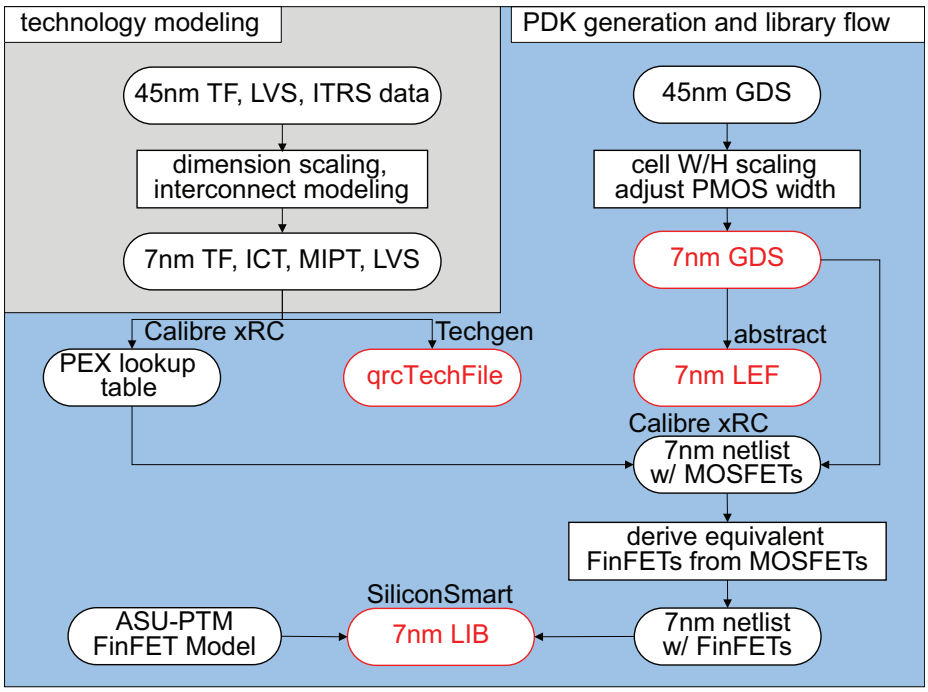

Figure 2.7: Predictive $7nm$ FinFET PDK and cell library generation flow based on NanGate FreePDK45 Open Cell Library.

Cu atoms from diffusing into and contaminating the dielectric layer. The resistivity of Cu is based on ITRS projections while the resistivity of TaN is determined to be $2000\mu\Omega \cdot cm$ [12]. Table 2.5 presents the resulting via resistance for each layers.

Contacts from M1 to active and poly utilize Tungsten (W) instead of Cu because of its excellent step coverage and gap fill abilities, especially for high-aspect ratio fills. Additionally, Tungsten Silicide (WSi_2) allows for low resistance contacts to the transistors. The resistivity of W contacts is determined to be $30\mu\Omega \cdot cm$ as projected in [13], which yields 27.3Ω and 46.14Ω for the resistance of active-M1 contacts and poly-M1 contacts, respectively.

7nm Standard Cell Library

Standard Cell Layout Scaling Ever since the introduction of multiple patterning for min-pitch metals in sub-$20nm$ nodes, W local interconnects (also called middle-of-line (MOL) layers) are used for cell-level routing. Since standard cell layouts with these fea-

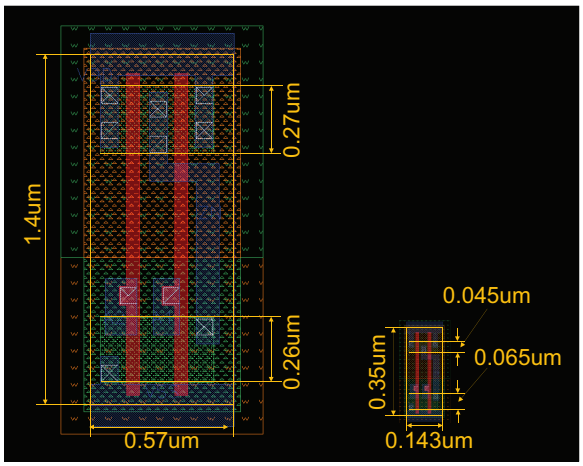

Figure 2.8: Comparison of NAND cell layouts between (a) NanGate FreePDK45 Open Cell Library and (b) the $7nm$ cell library

tures are not available publicly, the $45nm$ layouts are scaled down to $7nm$ dimensions, but MOL layers are not modeled in this work. This will result in some optimism when estimating cell-level parasitics, but the larger scope of this work remains unaffected because important parameters such as transistor behavior and interconnect parasitics are accurately modeled. The goal of this work is to understand important trends and trade-offs when working with future technologies.

The cell widths and heights of NanGate FreePDK45 Open Cell Library were shrunk along the xy-dimension with the scaling factor derived in process modeling (i.e., 0.25). For planar MOSFETs, electron mobility is higher compared to holes, and hence, PMOS transistors are sized wider. In sub-$45nm$ technologies, strain engineering improves carrier mobility and has been an important knob to improve performance every technology node. Additionally, PMOS transistors benefit more from strain resulting in nearly equal current drive-strengths as NMOS [14]. Hence, after scaling the $45nm$ planar layouts, the PMOS are sized equal to NMOS in order to balance cell rise and fall time.

An example $7nm$ NAND cell layout is compared with its $45nm$ counterpart in Figure 2.8. As shown in the figure, though cell height and width are scaled down according to the geometric scaling factor, PMOS width is shrunk further to balance drive-strength.

Library Exchange Format (LEF) views are created from $7nm$ cell layouts by Cadence® abstract generator. Interconnect dimensions and material properties discussed in previous subsections are coded in a MIPT file and is used to generate lookup tables for intra-cell parasitics using Mentor Graphics® Calibre® xRC™. These lookup tables, along with the scaled cell layouts and logic versus schematic (LVS) file, are used to extract intra-cell parasitics using Mentor Graphics® Calibre® xRC™.

Planar Width to Quantized Fins Since the $7nm$ cell layouts are scaled from the $45nm$ layouts assuming planar MOSFETs, the device widths have to be appropriately quantized to fins. The maximum number of fins in a standard cell is determined by the standard cell height and the ratio between metal pitch and fin pitch. Dummy fins are assumed in the cell layouts, which are required to make room for gate contacts between the FETs and to allow isolation between the FETs in adjacent cell rows. Therefore, the number of fins in a PMOS and NMOS pair is limited by the number of M1 tracks subtracted by the number of dummy fins. As Table 2.5 shows, the scaled design has ten M1 tracks, and a fin-pitch of $25.5nm$ is assumed to fit four fins per FET, which is in line with industry trends [10]. With an assumption of two dummy fins per PMOS and NMOS pair, dividing the transistor width by fin-pitch provides us the number of fins for that device.

Table 2.6 shows the maximum number of fins as well as the number of fingers derived using the method for various drive-strength cells of inverter. The low drive-strength inverters (i.e., INV_X1 to INV_X4) gain strength by increasing their number of fins while the high drive-strength inverters (i.e., INV_X8 to INV_X32) do so by increasing their number of fingers.

With intra-cell parasitic, quantized fins, and ASU PTM-MG FinFET transistor models for both HP and LSTP applications [15], timing and power metric (i.e., Liberty file) is generated using Synopsys SiliconSmart®.

Table 2.6: Maximum number of fins and the finger count in various drive-strength inverters in the $7nm$ cell libraries

cell name	max. fin count	finger count
INV_X1	1	1
INV_X2	2	1
INV_X4	4	1
INV_X8	4	2
INV_X16	4	4
INV_X32	4	7

Resulting Cell Libraries and Validation $7nm$ HP and LSTP cell libraries are generated with total 122 cells. Table 2.7 shows the cell delay, internal power-delay product (PDP) and leakage power of five selected cells comparing NanGate FreePDK45 Open Cell Library, the $7nm$ HP and $7nm$ LSTP cell libraries[2]. Figure 2.9 also shows the I-V characteristics of the transistor models used in cell characterization.

Compared to NanGate FreePDK45 Open Cell Library, the $7nm$ HP cell library has 84.7% lower cell delay on average. Due to the decrease in cell delay, voltage scaling and smaller input capacitance, the internal PDP of the $7nm$ HP cell library is significantly smaller. The leakage power consumption of the $7nm$ HP cell library are also 69.5% smaller on average, mainly due to the reduced dimension and supply voltage even though the $7nm$ HP transistor model shows higher I_{off}. The $7nm$ LSTP cell library has longer cell delay compared to the $7nm$ HP cell library because of the smaller I_{on} of lower leakage transistors as shown in Figure 2.9, but the internal PDP of the LSTP cell library is lower than the HP cell library. Since the $7nm$ LSTP transistor model is designed to specifically reduce its leakage power consumption, it exhibits much smaller I_{off} than the $45nm$ transistor model, which does not take leakage power reduction into account in its design.

Figure 2.10 shows the comparison of the 10-stage fanout-1 (FO1) INV delay between the projected values in [16] and the $7nm$ cell. The $7nm$ INV cell delay is within 10% of the projections made in [16]. Considering that both approaches utilize the same transistor

[2]In order to obtain a fair comparison between different technology nodes, the input slew is set to the output slew of INV_X4, and the output capacitance to the input capacitance of 4 INV_X4 cells of corresponding process.

Table 2.7: Timing and power metric comparison between NanGate FreePDK45 Open Cell Library, and the $7nm$ HP and LSTP cell libraries for five selected cells

cell name	$45nm$	$7nm$ HP		$7nm$ LSTP	
cell delay (ps)					
BUF_X4	44.4	7.3	(-83.5%)	15.4	(-65.4%)
DFF_X2	114.9	15.9	(-86.2%)	33.4	(-70.9%)
INV_X4	21.4	4.1	(-80.8%)	8.3	(-61.4%)
NAND2_X2	38.8	6.5	(-83.3%)	12.6	(-67.5%)
NOR2_X2	46.1	6.6	(-85.6%)	12.9	(-72.0%)
internal PDP (fJ)					
BUF_X4	16.25	0.186	(-98.9%)	0.156	(-99.0%)
DFF_X2	7.25	0.430	(-94.1%)	0.396	(-94.5%)
INV_X4	5.97	0.103	(-98.3%)	0.084	(-98.6%)
NAND2_X2	1.78	0.080	(-95.5%)	0.066	(-96.3%)
NOR2_X2	4.48	0.090	(-98.0%)	0.078	(-98.3%)
leakage power (nW)					
BUF_X4	41.5	13.8	(-66.7%)	0.049	(-99.9%)
DFF_X2	200.6	41.2	(-79.5%)	0.139	(-99.9%)
INV_X4	40.3	11.1	(-72.4%)	0.012	(-100.0%)
NAND2_X2	29.6	10.7	(-63.8%)	0.012	(-100.0%)
NOR2_X2	42.4	11.1	(-73.8%)	0.014	(-100.0%)

models, the plot shows the accuracy of the cell-level parasitics and hence, the efficacy of the predictive $7nm$ FinFET PDK and cell libraries.

2.3.2 Power Benefits of Monolithic 3D ICs in $7nm$ Technology Nodes

Experimental Setup

2D and M3D ICs are implemented with maximum achievable clock frequencies for AES-128, fast fourier transform (FFT), and LDPC from OpenCores using the developed $7nm$ PDK and cell libraries. The footprint of each design is determined by targeting cell utilization to be 60% for AES-128 and FFT, and to 40% for LDPC since LDPC is a wire-dominated circuit, so that the chip area is determined not by cell utilization, but the available routing resources. Statistical power simulation is used to derive power metrics of the implemented designs, and considering dark-silicon in the advanced technology nodes [17], only 30% of sequential logics (i.e., flip-flops) are assumed to be powered-on at an instance.

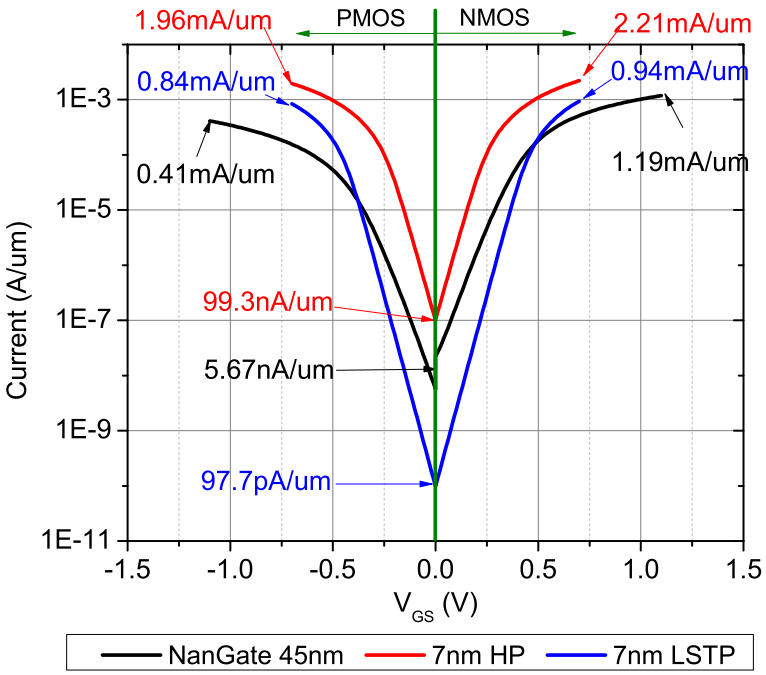

Figure 2.9: Comparison of I_{on} and I_{off} for the unit-width $45nm$, $7nm$ HP and LSTP transistor models. Values for $7nm$ transistor models are derived by measuring current flowing the transistor with single fin and normalizing the current by effective width ($2\times H_{FIN}\times T_{FIN}$).

The toggle rate for sequential logics and primary inputs are set to 40% and 20%, respectively. Table 2.8 and Table 2.9 show the design and power metrics of the three benchmarks comparing the M3D ICs to their 2D counterparts, respectively.

Power Saving Results

To interpret the trends in Table 2.8 and Table 2.9, the following equation is deployed which describes the components of the dynamic power consumption of a design.

$$P_{dyn} = P_{INT} + \alpha \cdot (C_{pin} + C_{wire}) \cdot {V_{DD}}^2 \cdot f_{clk}, \tag{2.8}$$

where P_{INT}, is the cell internal power, and the second term describes net switching power where C_{pin} and C_{wire} are the pin and wire capacitance in the design.

M3D ICs offer power benefit in two ways. The first component comes from the reduced C_{wire} in the second term of Equation (2.8) due to wire-length reduction. Since an M3D IC

Table 2.8: Iso-performance design metric comparison between 2D and M3D ICs with $7nm$ HP and LSTP cell libraries. The percentage values in M3D ICs are computed with respect to their 2D counterparts.

		$7nm$ HP			$7nm$ LSTP		
design	parameter	2D	M3D		2D	M3D	
AES-128	clock frequency (MHz)	12,000	12,000	(0.0%)	6,000	6,000	(0.0%)
	footprint (μm)	140×139	101×101	(47.6%)	140×140	99×99	(49.9%)
	cell area (μm^2)	13,353	12,564	(5.9%)	14,358	12,930	(9.9%)
	wire-length (μm)	539,859	464,552	(13.9%)	479,148	442,215	(7.7%)
	avg. net size	3.008	2.790	(7.2%)	3.022	2.854	(5.6%)
	MIV count	-	47,370		-	46,676	
LDPC	clock frequency (MHz)	2,500	2,500	(0.0%)	1,200	1,200	(0.0%)
	footprint (μm)	96×95	68×68	(49.8%)	96×95	68×68	(49.5%)
	cell area (μm^2)	4,130	3,970	(3.9%)	4,258	3,787	(11.1%)
	wire-length (μm)	630,826	493,876	(21.7%)	617,071	464,610	(24.7%)
	avg. net size	3.402	3.028	(11.0%)	3.356	3.089	(8.0%)
	MIV count	-	22,067		-	22,913	
FFT	clock frequency (MHz)	5,000	5,000	(0.0%)	2,500	2,500	(0.0%)
	footprint (μm)	245×245	173×174	(49.9%)	245×245	173×173	(50.0%)
	cell area (μm^2)	39,452	37,156	(5.8%)	37,591	36,463	(3.0%)
	wire-length (μm)	1,424,382	1,159,516	(18.6%)	1,361,842	1,036,165	(23.9%)
	avg. net size	3.164	3.129	(1.1%)	3.247	3.190	(1.8%)
	MIV count	-	85,379		-	78,476	

Table 2.9: Iso-performance power metric comparison between 2D and M3D ICs with $7nm$ HP and LSTP cell libraries. The percentage values in M3D ICs are computed with respect to their 2D counterparts.

design	parameter	$7nm$ HP			$7nm$ LSTP		
		2D	M3D		2D	M3D	
AES-128	clock frequency (MHz)	12,000	12,000	(0.0%)	6,000	6,000	(0.0%)
	cell internal power (mW)	87.5	76.1	(13.0%)	29.4	27.0	(8.2%)
	net switching power (mW)	73.4	64.9	(11.6%)	47.1	40.8	(13.4%)
	leakage power (mW)	2.148	1.178	(45.2%)	0.004	0.004	(0.0%)
	total power (mW)	**162.8**	**142.2**	**(12.7%)**	**76.5**	**67.8**	**(11.4%)**
LDPC	clock frequency (MHz)	2,500	2,500	(0.0%)	1,200	1,200	(0.0%)
	cell internal power (mW)	10.8	9.2	(14.8%)	2.8	2.4	(14.3%)
	net switching power (mW)	46.2	32.7	(29.2%)	20.4	14.2	(30.4%)
	leakage power (mW)	0.382	0.336	(12.0%)	0.001	0.001	(0.0%)
	total power (mW)	**57.4**	**42.2**	**(26.5%)**	**23.2**	**16.6**	**(28.4%)**
FFT	clock frequency (MHz)	5,000	5,000	(0.0%)	2,500	2,500	(0.0%)
	cell internal power (mW)	139.7	130.4	(6.7%)	37.4	34.8	(7.0%)
	net switching power (mW)	75.3	65.7	(12.7%)	31.6	27.6	(12.7%)
	leakage power (mW)	4.122	3.835	(7.0%)	0.013	0.012	(7.7%)
	total power (mW)	**219.1**	**199.9**	**(8.8%)**	**69.0**	**62.4**	**(9.6%)**

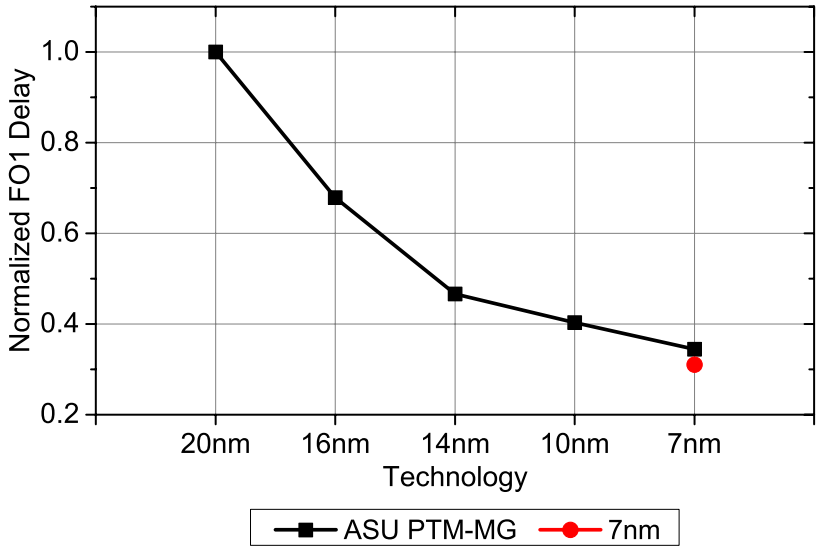

Figure 2.10: Normalized FO1 cell delay of a 10-stage INV_X4 chain

utilizes short vertical connections (i.e., MIVs) instead of long horizontal wires in the xy-plane, it reduces the wire-length of the design. Since wire-length is closely related to wire capacitance, C_{wire}, it helps reduce the net switching power of the design.

The second component is attributed to its standard cell area saving. The reduced C_{wire} help an M3D IC meet timing more easily, utilizing less number of buffers and lower drive-strength cells and hence, resulting in standard cell area reduction. Figure 2.11 shows the cell drive-strength distribution of FFT $7nm$ HP 2D and M3D ICs. It is evident that the M3D IC uses smaller cell sizes, utilizing more cells with lower drive-strength. The standard cell area saving helps reduce both the first term, P_{INT}, and the second term by reducing C_{pin} in Equation (2.8).

Impact of Characteristic of Benchmarks

There is a significant difference in the total power saving between LDPC and the other benchmarks. While the total power saving of the AES-128 and FFT M3D ICs ranges from 8.8% to 12.7%, the LDPC M3D ICs show 26.5% to 28.4% power saving. To explain the difference, Equation (2.8) is re-written as follows:

$$P_{dyn} = P_{INT} + \alpha \cdot C_{pin} \cdot {V_{DD}}^2 \cdot f_{clk} + \alpha \cdot C_{wire} \cdot {V_{DD}}^2 \cdot f_{clk} \tag{2.9}$$

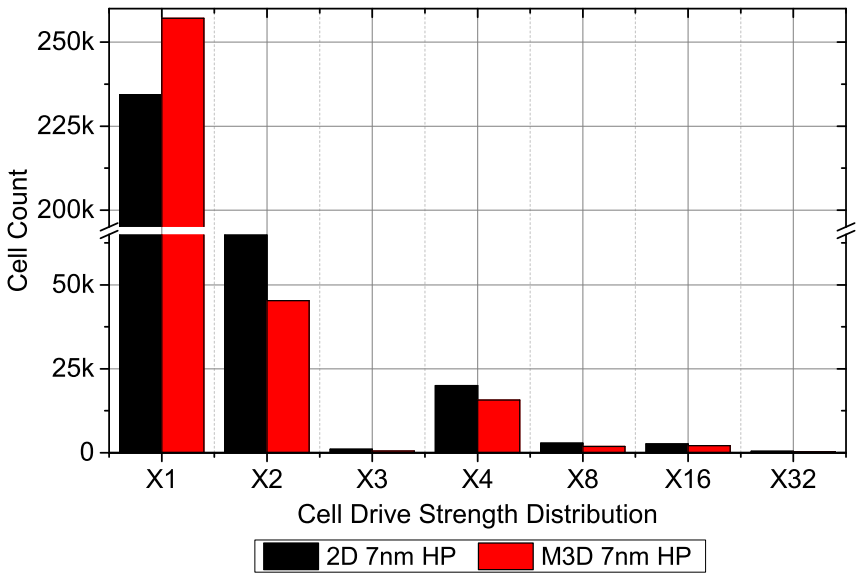

Figure 2.11: Cell drive-strength distribution of FFT $7nm$ HP 2D and M3D ICs

First, the ratio of the net switching power to the cell internal power of LDPC $7nm$ HP 2D IC (= 4.27) is much higher than the other two designs (AES-128 = 0.84, FFT = 0.54) in Table 2.9, which means the first term is much smaller than the other two terms in Equation (2.9) in the LDPC design.

In addition, from Table 2.8, the ratio of the wire-length to the standard cell area of the LDPC design is 150.74μm^{-1}, which is also much higher than the AES-128 (= 40.43μm^{-1}) and FFT (= 36.10μm^{-1}) designs. Hence, the ratio of C_{wire} to C_{pin} is much larger in the LDPC design than that in the other two benchmarks since C_{pin} and C_{wire} is highly correlated to standard cell area and wire-length of a design, respectively.

Considering that the standard cell area saving reduces the first two terms in Equation (2.9), and the wire-length saving reduces the last term, the total power saving of the LDPC design heavily depends on the wire-length saving of the designs rather than the standard cell area saving.

Moreover, M3D stacking technology achieves more wire-length saving in the LDPC design than the other two benchmarks. The ratio of the wire-length to the footprint in the LDPC $7nm$ HP 2D IC (= 69.17μm^{-1}), is also larger than the other two designs (AES-128

= 27.74μm^{-1}, FFT = 23.73μm^{-1}). This metric along with the average net size in Table 2.8 indicate that wires are more congested in the LDPC design. Since an M3D IC helps reduce the wire congestion by utilizing both top and bottom metal layers, it effectively reduces the wire-length in the LDPC design.

Impact of Cell Library

The trend difference between the $7nm$ HP and $7nm$ LSTP designs can be also explained by the ratio of C_{wire} to C_{pin} in Equation (2.9). Note that the wire RC parasitics of both libraries are remained same because they share the same cell layouts. Considering the fact that the pin capacitance of the $7nm$ HP cells tend to be higher than that of the $7nm$ LSTP cells due to higher parasitics in HP devices, the ratio of the third term in Equation (2.9) to the total dynamic power is larger in the $7nm$ LSTP designs than in the $7nm$ HP designs. This indicates that designs with the LSTP cell library take more advantage from the wire-length saving of M3D stacking technology than those using the HP cell library. Since M3D ICs reduce wire-length more effectively in the LDPC design compared to the other two benchmarks, the largest total power saving is offered in the LDPC $7nm$ LSTP M3D ICs than the others.

2.3.3 Computer-Aided Design Solutions to Improve Power Saving of Monolithic 3D ICs

Since the design quality of an M3D IC heavily depends on the tier partitioning methodology of the M3D IC design flow described in Section 2.2.1, tier partitioning schemes are investigated to improve the power saving of M3D ICs.

Bin Size Selection

As described in Section 2.2.1, a *shrunk-2D design* is, first, divided into multiple square bins on the xy-plane, and the area-balanced min-cut partitioning algorithm is performed for each bin.

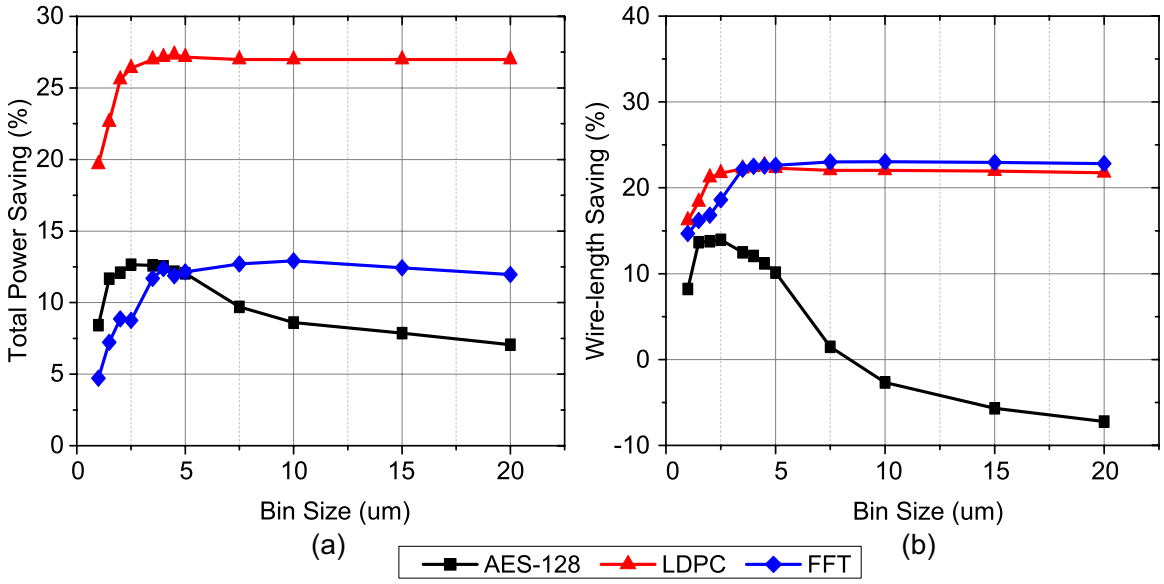

Figure 2.12: Impact of bin size selection on (a) the total power saving and (b) wire-length saving of M3D ICs implemented using the $7nm$ HP cell library

The impact of the bin size on the total power saving of M3D ICs over 2D ICs is shown in Figure 2.12 (a). For each benchmark, twelve M3D ICs are implemented from a single *shrunk-2D design* using different bin sizes, and the power consumption of the resulting M3D ICs are compared with the corresponding 2D ICs. The total power saving of all 3 benchmarks are maximum at the bin size of 3.0$\sim$4.0μm, and the total power saving is ranged from 8.4% to 12.7% for the AES-128 M3D ICs. Another important trend to note is that unlike the LDPC and FFT M3D ICs, the total power savings of the AES-128 M3D ICs sharply decrease after 5.0μm bin size.

Considering that the cell count and drive-strength of cells are not changed during the tier partitioning step, the difference on the total power saving comes mainly from the difference on the wire-length saving due to the bin size selection, which shows similar trend as shown in Figure 2.12 (b).

Figure 2.13 explains the reason why the bin size selection affects the wire-length saving of an M3D IC. If the bin size during tier partitioning is more than 5.0μm (i.e., large bin size), the area-balanced min-cut partitioning algorithm finds global optimal solution more easily, minimizing the number of vertical connections between two tiers. However, neigh-

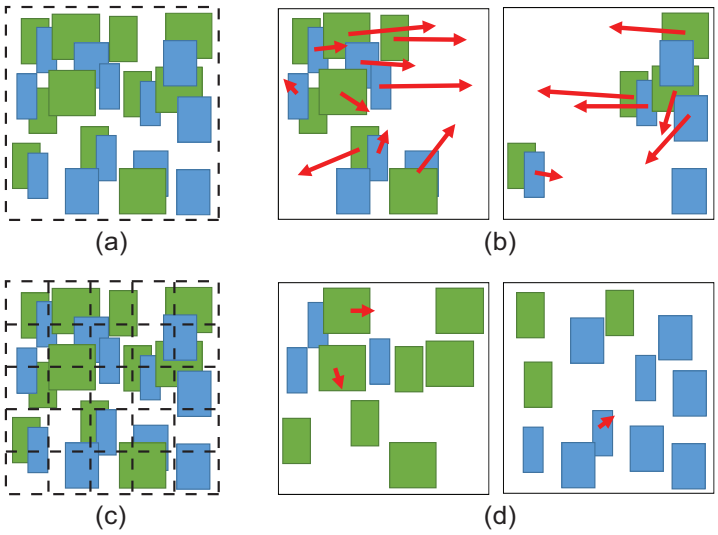

Figure 2.13: Illustration showing how bin size selection affects the wire-length saving of M3D ICs. (a) *Shrunk-2D design* with only one large bin (b) and the corresponding partitioning result. (c) *Shrunk-2D design* with very small sized bins (d) and the corresponding partition result. Dashed-lines and red arrows indicate bins and cell movement during legalization, respectively.

boring cells in local area tend to be clustered and placed on a single tier altogether, leaving large overlap among cells even after the tier partitioning. This overlap increases cell movement while legalizing cells (red arrows in the Figure 2.13 (b)), resulting in wire-length overhead. The local cell clustering due to large bin size becomes more severe if a design has clustered design structure as AES-128 designs in Figure 2.14 (a), which indicates that the design has the large number of local wires, but very few global wires.

On the other hand, in the case of bin size less than $3.0\mu m$ (i.e., small bin size), it does not suffer from overlap after tier partitioning because of its fine-grained partitioning scheme as shown in Figure 2.13 (c) and Figure 2.13 (d), but it is more likely to fall into local optimal solution during tier partitioning, splitting unnecessarily large number of local interconnects into two tiers, showing the sharp decrease as shown in Figure 2.12 (b).

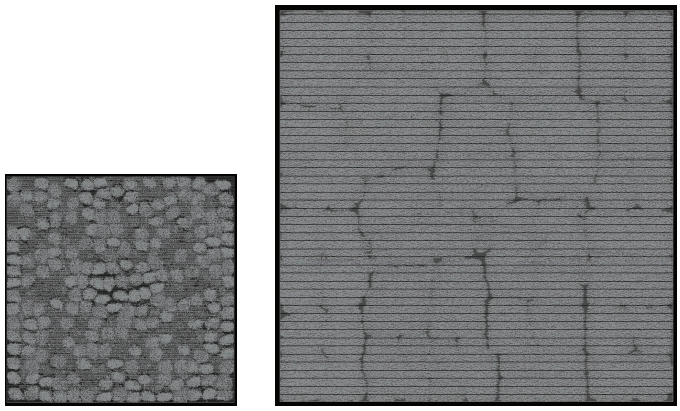

Figure 2.14: Comparison of cell placement of (a) AES-128 and (b) FFT *shrunk-2D designs*. AES-128 has clustered design structure while FFT has evenly distributed cell placement.

Clock Tree Partitioning Methodology

Two clock tree partitioning methodologies, level of freedom (LOF)-based partitioning [9] and prioritized clock tree partitioning, to improve the clock tree of M3D ICs are presented.

Level of Freedom-based Partitioning The LOF of a clock tree is defined as the distance from the leaf nodes of the clock tree, and the clock cells within the distance are free to be partitioned either the top or bottom tier as shown in Figure 2.15. All other cells whose distance is larger than the LOF are fixed on one of the tiers (e.g., the top tier) before partitioning. For example, if LOF = 1, only leaf cells of a clock tree (i.e., flip-flops), and the clock buffers which drive the leaf cells are free to be partitioned on either tiers, and every other clock cells are fixed on the top tier before the tier partitioning.

Table 2.10 shows the number of clock cells fixed to the top tier before tier partitioning depending on LOF and the resulting number of clock MIVs. Since clock MIVs are used when a parent cell in a clock tree is assigned to the different tier from its child cells, the number of clock MIVs are increasing as LOF increases. Figure 2.16 compares a clock trees on the top tier which span from their root to the first MIV encountered along the branches when LOF = 1 and LOF = max. The figure clearly shows that the clock tree spanning from

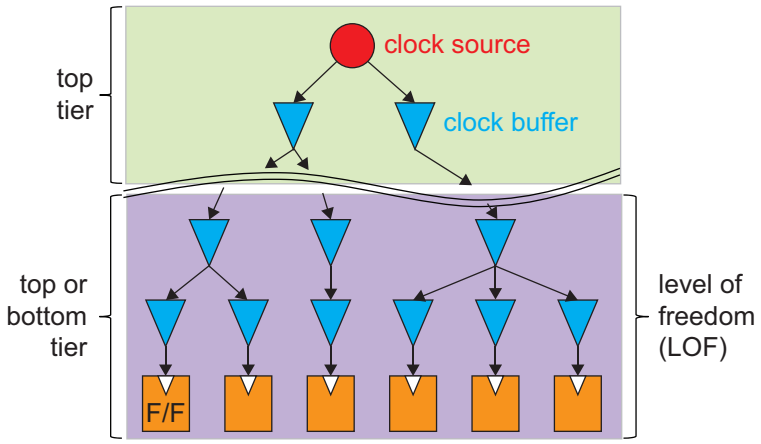

Figure 2.15: An example showing the concept of LOF in a clock tree when LOF = 2. Note that the root of a clock tree, the clock source, is always placed on the top tier.

Table 2.10: Number of the fixed clock cells and the resulting clock MIVs in M3D ICs implemented using the $7nm$ HP cell library depending on LOF.

LOF	AES-128		LDPC		FFT	
	fixed clk cell count	clk MIV count	fixed clk cell count	clk MIV count	fixed clk cell count	clk MIV count
0	1,057	379	239	64	1,644	1,498
1	732	390	176	67	121	1,522
2	497	385	128	65	61	1,510
3	305	402	83	78	42	1,507
4	162	411	65	71	30	1,503
5	77	403	49	79	21	1,506
max	0	413	0	246	0	5,485

the clock source is larger as LOF decreases, since more number of clock cells are fixed on to the top tier with low LOF.

Figure 2.17 (a) presents the impact of LOF on the clock switching power saving of M3D ICs over the corresponding 2D ICs, showing that the clock switching power saving decreases as LOF increases. However, the total power consumption of the designs does not vary by a large magnitude, since the clock power consumption is less than 3% of the total power consumption for all benchmarks. On the other hand, the clock skew of the designs is significantly affected by LOF as shown in Figure 2.17 (b).

The decreased clock switching power saving and the increased clock skew in M3D ICs with high LOF are attributed to the increased clock wire-length of the designs. As

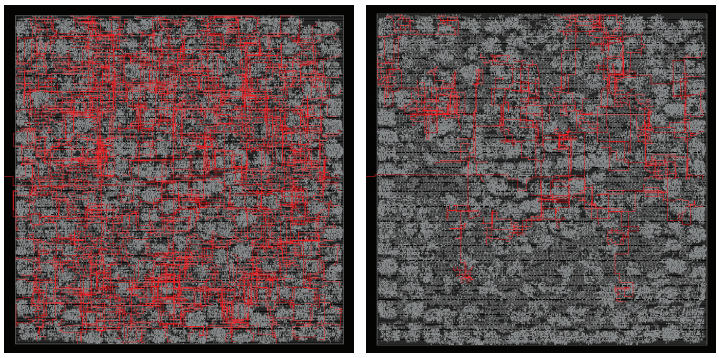

Figure 2.16: Clock trees on the top tier span from the root to the first clock MIV encountered in the AES-128 $7nm$ HP M3D IC when (a) LOF = 1 and (b) LOF = max

LOF affects only on the tier assignment of clock cells (i.e., the z-location), not on the xy-location of the cells, if LOF increases, the number of connection crossing the tiers is increasing, which, in turn, worsens the clock skew of the M3D ICs as discussed in Figure 2.4. In addition, they also increase the RC parasitics of the clock tree, increasing the clock switching power consumption of the design.

Prioritized Clock Tree Partitioning In this technique, the tier partitioning stage is divided into two phases, clock cell pre-tier partitioning phase and regular cell tier partitioning phase as shown in Figure 2.18. The entire clock tree of a design, which includes clock buffers as well as flip-flops, is first partitioned into two tiers using the area-balanced min-cut partitioning algorithm. The partitioned clock cells are fixed on the assigned tier and are not changed thereafter. Then, regular cells are partitioned, so that the cell area of the top and bottom tier are balanced.

Different from LOF-based partitioning with LOF = max, which partitions all cells at the same time, the methodology gives clock cells higher priority over regular cells while balancing the area of clock cells on the top and the bottom tier.

As discussed in Section 2.3.3, in order to minimize the clock skew and the clock switching power of an M3D IC, it is important to reduce the number of connection between clock cells crossing tiers. Therefore, in clock cell pre-tier partitioning phase, larger bin size is

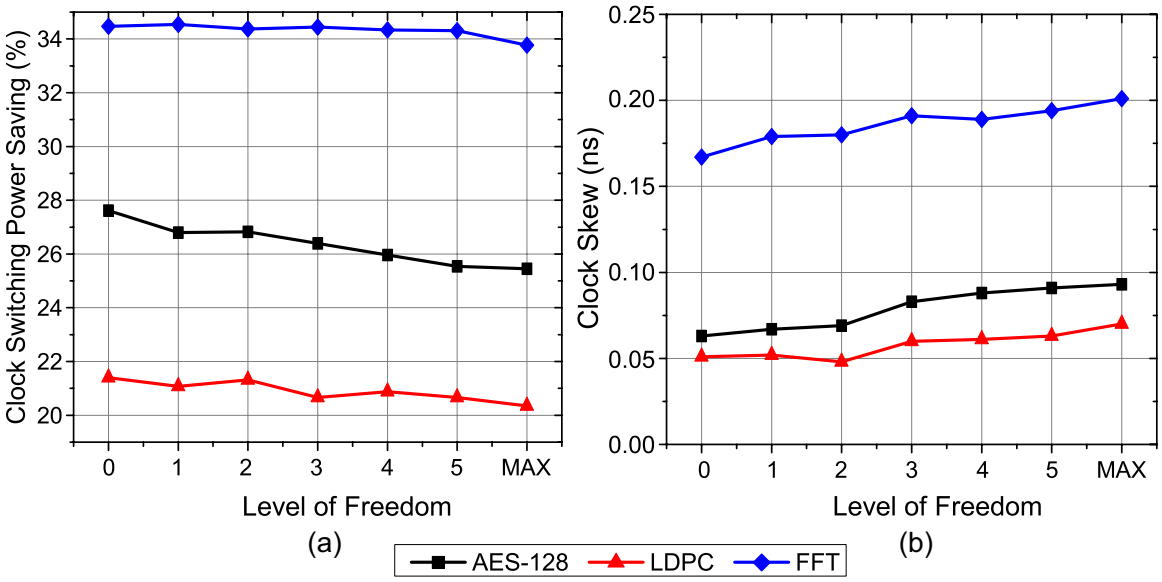

Figure 2.17: Impact of LOF on (a) the clock switching power and (b) clock skew of $7nm$ HP M3D ICs

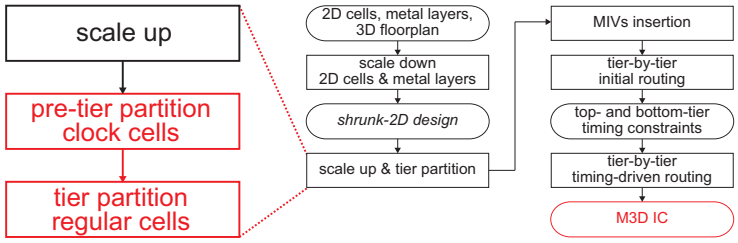

Figure 2.18: M3D IC design flow with prioritized clock tree partitioning extended from Figure 2.1

utilized, compared to the bin size which would be used for regular cells, to minimize the cut size of the clock tree during the area-balanced min-cut partitioning algorithm. The larger bin size may cluster neighboring clock cells in local area and place all of them onto one tier as shown in Figure 2.13 (b), but considering that the number of clock cells are small compared to the total cell count of a design, the issue is automatically resolved during partitioning regular cells in the next phase by placing more regular cells on the other tiers. In this work, the bin size for clock cell pre-tier partitioning is determined to be maximum in the range that regular cell tier partition is able to balance the standard cell area between two tiers.

Table 2.11 shows the impact of the prioritized clock tree partitioning technique on the

Table 2.11: Impact of the prioritized clock tree partitioning technique on the design and power metric of clock trees in M3D ICs. LOF-based partitioning with LOF = 0 is used for the baseline.

design	parameter	baseline	prioritized clock partitioning	
AES-128	bin size (μm)	-	101	
	clock MIV count	379	118	(-68.9%)
	clk skew (ns)	0.063	0.058	(-7.9%)
	clk sw power (mW)	0.321	0.307	(-4.4%)
LDPC	bin size (μm)	-	68	
	clock MIV count	64	26	(-59.4%)
	clk skew (ns)	0.051	0.043	(-15.7%)
	clk sw power (mW)	0.072	0.070	(-2.4%)
FFT	bin size (μm)	-	21.75	
	clock MIV count	1,498	1,220	(-18.6%)
	clk skew (ns)	0.167	0.138	(-17.4%)
	clk sw power (mW)	1.954	1.942	(-0.6%)

design and power metric of the clock trees in M3D ICs. The methodology successfully reduces the number of clock MIVs by finding the global optimal solution with large bin size. The reduced clock MIVs minimize the inter-tier routing overheads, achieving 2.5% clock switching power reduction and 13.7% clock skew reduction on average. An interesting point to note is that while the clock switching power saving is proportional to the clock MIV reduction because of the reduced inter-tier routing overhead, the benefit on the clock skew does not because the reduced clock MIVs also reduce the clock latency of the shortest branch of the clock tree as well as the longest branch although it benefits more on the longest branch.

2.4 Summary

2.4.1 Monolithic 3D IC Performance Improvement

The observations and guidelines to improve the performance of M3D ICs are summarized as follows:

- M3D ICs offer performance improvement over 2D ICs, showing up to 15.6% higher

maximum achievable clock frequency

- The parasitic adjustment method helps estimate the RC parasitics of M3D ICs more accurately and insert minimal buffers in the right place. The method presents benefits not only in performance, but also power consumption of M3D ICs.
- Pre-placing cells in a number of critical timing paths in a single tier helps improve the performance of M3D ICs by reducing the inter-tier routing overhead in critical timing paths
- Partitioning clock buffers into a single tier minimizes the number of MIVs in a clock tree. This reduces the clock skew of timing paths, offering increased timing budget in sequential cell dominated circuits.

2.4.2 Power Saving of Monolithic 3D ICs in Advanced Technology Nodes

The findings when adopting M3D stacking technology in $7nm$ technology nodes for low power applications are summarized as follows:

- M3D stacking technology offers significant power saving at a $7nm$ technology node, convincingly showing that M3D ICs offer consistent power saving across device generations and target applications
- Designs with LSTP devices benefits more from M3D ICs than HP devices because the ratio of wire capacitance to total capacitance is larger in LSTP designs
- Bin size during tier partitioning significantly affects the total power saving. While small bin size is prone to fall into local optimal solution during the area-balanced min-cut partitioning, large bin size increases wire-length during legalization of cells after partitioning especially for designs with clustered design structure.
- Although clock tree partitioning methodology does not have large impact on the total power saving of M3D ICs at advanced technology nodes, it affects the clock skew of

designs. With LOF-based clock tree partitioning, in order to reduce the clock skew of M3D ICs, fixing all clock buffers on a tier is recommended.

- Prioritized clock tree partitioning improves the quality of the clock tree of a design by giving clock cells higher priority over regular cells, which, in turn, reduces the inter-tier routing overheads during the clock tree design of M3D ICs

CHAPTER 3

NEW MONOLITHIC 3D IC DESIGN FLOW

3.1 Motivation and Background

The industry has transitioned from planar MOSFETs to 3D FinFETs at the 14/16nm node to combat worsening electrostatics and degraded short channel effects due to channel length scaling. Improved transistor characteristics in FinFETs are achieved at the cost of higher parasitic capacitance associated with the 3D fins and the introduction of the local interconnects that are needed to contact the devices to metal routing layers. Due to limited viable transistor options beyond FinFETs and the increasing cost and complexity of lithography strategies to print sub-7nm node features, traditional Moore's law scaling is slowing down. These limitations create a technology inflection point for 'More than Moore' technologies [18] such as M3D ICs to bring value and be adopted into mainstream designs.

Although previous studies have explored different aspects of M3D ICs and shown improvement in power and performance [6], most, if not all of these studies have used open-source academic benchmarks and libraries to demonstrate the effectiveness of M3D stacking technology. In spite of being a valuable research tool, academic benchmarks and libraries have limited validity and accuracy when compared to commercial designs.

In order to be deployed in real-word designs, M3D ICs need to be cost-effective and deliver power or performance improvement of the order of magnitude similar to that obtained by Moore's law process scaling. Evaluating cost-effectiveness is non-trivial as M3D stacking technology is still under active research and development. Hence, the power improvement of M3D ICs in a real-world design - an in-order, 32-bit application processor - is evaluated while assessing whether or not that improvement is independent of the underlying technology node.

Currently EDA tools do not support M3D ICs, and hence, previous studies have explored implementation approaches of M3D ICs using 2D commercial tools. In [5], in order to estimate cell placement and wire-length of an M3D IC, the dimensions of cells and wires are shrunk, and a *shrunk-2D design* is implemented in half footprint of the 2D IC. However, using *shrunk-2D designs* are prone to inaccurate buffer insertion because of inaccurate wire-load estimation as discussed in Section 2.2.1. Moreover, the flow is completely design-agnostic, utilizes very large number of MIVs and hence, partitions local cells into separate tiers resulting in a non-optimal tier partition. Another M3D IC design flow is proposed in [19], which folds 2D placement at the center of the die into two separate tiers. However, using their design flow shows marginal wire-length savings and no power savings, and does not take into account design details to guide partitioning, resulting in a non-optimal solution. Therefore, a new M3D IC design flow is necessitated which incorporates design and micro-architecture information during partitioning cells on multiple tiers while supporting accurate buffer insertion with accurate wire-load estimation.

3.2 Benefit Trends of Monolithic 3D ICs across Technology Nodes

In this work, a comprehensive study investigating the power impact of M3D ICs across technology nodes is presented using a commercial in-order 32 bit application processor on foundry $28nm$, foundry 14/16nm and predictive $7nm$ technology nodes. Based on the observation of the work, M3D stacking technology provides maximum power savings at the $28nm$ technology node, and the benefits improve at higher clock frequencies with the reduction of standard cell area in addition to wire-length savings. An in-depth analysis of the results and guidelines for M3D ICs are presented to support the observations in this work.

Table 3.1: Key metrics of foundry $28nm$, 14/16nm and the predictive $7nm$ technology nodes used in this work

parameters	$28nm$	14/16nm	$7nm$
transistor type	planar	FinFET	FinFET
V_{DD} (V)	0.9	0.8	0.7
CPP (nm)	110∼120	78∼90	50
M1 pitch (nm)	90	64	36
MIV cross-section (nm)	80×80	40×40	32×32
MIV height (nm)	140	170	170

3.2.1 Analysis on Benefits of Monolithic 3D ICs

Technology Nodes and Design Libraries

Table 3.1 shows the representative metrics of each technology nodes used in this work, based on previous publications [20, 14, 21, 22]. The $28nm$ technology node is planar transistor based while 14/16nm is the first generation foundry FinFET technology node. For these technology nodes, production-level standard cell libraries are utilized containing over 1,000 cells and memory macros that were designed, verified, and characterized using foundry PDK.

Since $7nm$ technology node parameters are still under development by foundries, a more realistic predictive PDK, which is different from the one described in Section 2.3.1, is developed to generate the required views for this work. The predictive $7nm$ PDK contains electrical models (BSIM-CMG), DRC, LVS, extraction and LEF files. The transistor models incorporate scaled channel lengths and fin-pitches and increased fin-heights compared to previous technology nodes in order to improve performance at lower supply voltages. Multiple threshold voltages (V_T) and variation corners are supported in the predictive $7nm$ PDK. Process metrics such as gate pitch and metal pitches are linearly scaled from previous technology nodes, and the design rules are created considering lithography challenges associated with printing these pitches. The interconnect stack is modeled based on similar scaling assumptions. Different from the standard cell libraries used in Section 2.3.1 which is scaled from NanGate FreePDK45 Open Cell Library, $7nm$ standard cell libraries and

memory macros are designed from the scratch and characterized using the PDK.

The M3D IC requires six metal layers on both top and bottom tiers. The MIVs connect M6 of the bottom tier with M1 of the top tier. The size of the MIVs is limited to be $2\times$ the minimum via size allowed in the technology node to reduce MIV resistance. The MIV heights take into account the fact that the MIVs need to traverse through inter-tier dielectrics and transistor substrates to contact to M1 on the top tier. The MIV height increases from $28nm$ to 14/16nm and $7nm$ technology nodes because of the introduction of local interconnect MOL layer in the sub-$20nm$ nodes.

Since M3D IC fabrication is done sequentially, high temperature front-end device processing of the top tier can adversely affect the interconnects in the bottom tier while low-temperature processing will result in inferior top tier transistors. Recent work reporting low-temperature processes that achieve similar device behavior across both tiers have been presented [23] and hence, all implementations in this work are done with the assumption of similar device characteristics in both tiers.

Implementation Methodology

The standard cell libraries and memory macros for the $28nm$, 14/16nm, and $7nm$ technology nodes are used to synthesize, place, and route the full-chip design. 2D and M3D ICs of the application processor are implemented sweeping the target frequency from $500MHz$ to $1.2GHz$ in $100MHz$ increments across the three technology nodes. M3D ICs are implemented using **shrunk-2D design flow** presented in Section 2.2.1. Full-chip timing is met at the appropriate corners (i.e., slow corner for setup and fast corner for hold). Power is reported at the typical corner. The floorplan of the design is customized for each technology node to meet timing, but kept constant during frequency sweeps. Multiple iterations of the 2D and M3D IC floorplan are required at each node to ensure that the designs meet timing. The chip area is fixed such that the final cell utilization is similar across technology nodes.

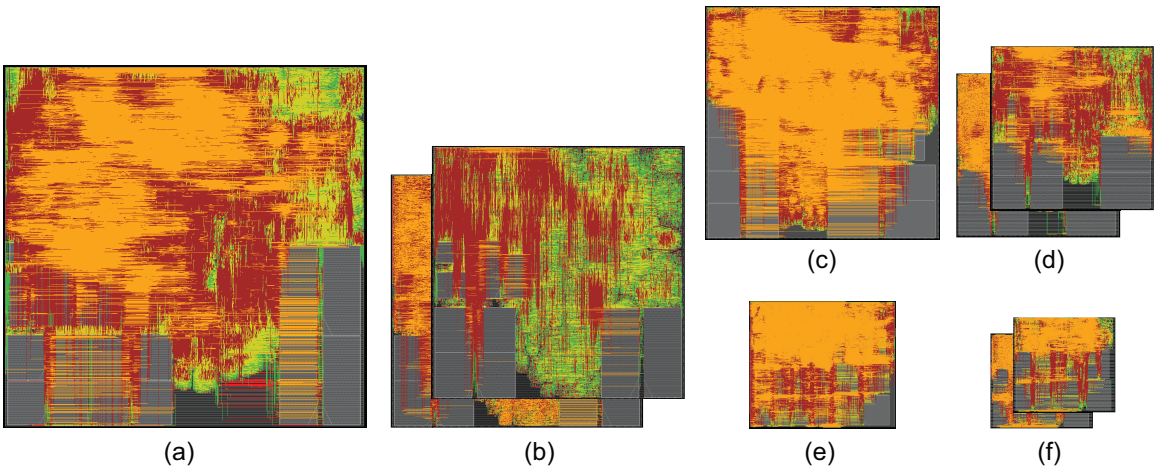

Figure 3.1: GDS layouts of (a) 28nm 2D, (b) 28nm M3D, (c) 14/16nm 2D, (d) 14/16nm M3D, (e) 7nm 2D, and (f) 7nm M3D ICs of the application processor at 1.1GHz.

Power Saving Trend of Monolithic 3D ICs

Figure 3.1 shows the GDS layouts of the 2D and M3D ICs of the application processor at 1.1GHz. The implementation tools are unable to meet timing at 1.2GHz target frequency for the 28nm and 14/16nm designs, hence their results are reported up to 1.1GHz, while the 7nm results up to 1.2GHz.

The normalized total power consumption of the 2D and M3D ICs across technologies are shown in Figure 3.2. The total power of both the 2D and M3D ICs increases with increasing frequency across technology nodes, which is expected. The power saving with the M3D ICs over the 2D ICs is shown in Figure 3.3. There are two important trends in Figure 3.3: (1) 28nm node shows the maximum power savings in M3D IC across all frequencies, and (2) the power saving of M3D stacking technology over 2D ICs (i.e., M3D power savings) increase with increasing target frequency of the designs.

To interpret and analyze the results, Equation (3.1) is used which is extended from Equation (2.8), and describes the components of dynamic power in an IC.

$$P_{dyn} = P_{INT} + \alpha \cdot (r_{p2w} \cdot C_{wire} + C_{wire}) \cdot {V_{DD}}^2 \cdot f_{clk}, \tag{3.1}$$

where r_{p2w} is the ratio of the pin capacitance to the wire capacitance. As discussed in

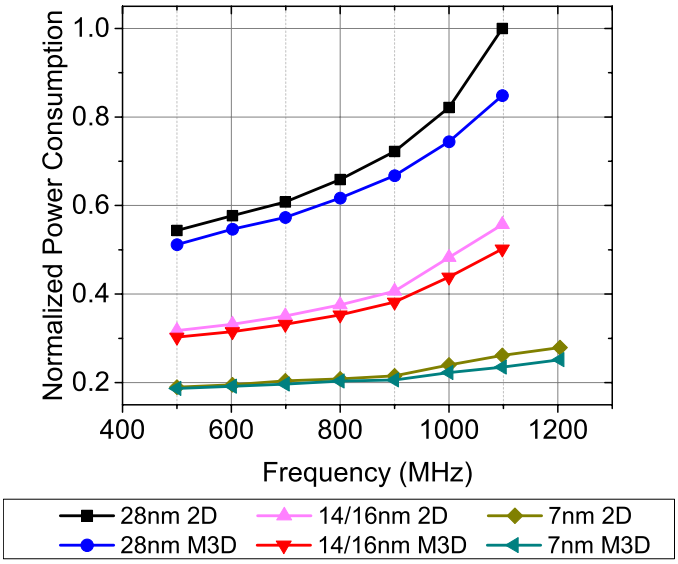

Figure 3.2: Normalized total power consumption of 2D and M3D ICs in 28nm, 14/16nm and 7nm technology nodes

Section 2.3.2, the primary advantage of M3D ICs comes from wire-length reduction resulting in reduced wire capacitance switching power dissipation. With the reduction in wires, the synthesis, place and route tools can also reduce the drive-strengths of the gates and buffers used to meet the design targets leading to reduced internal power (P_{INT}) and pin capacitance switching component as well. The total power reduction in an M3D IC depends on wire-length reduction, the number of cells and cell size reduction, the ratio of pin capacitance to wire capacitance and net switching power to internal power in the 2D IC.

Further extending Equation (3.1), as internal power and pin capacitance depends on standard cell area, and wire-length affects wire capacitance, M3D power savings can be denoted as follows:

$$\Delta P_{dyn} = \Delta_{cell} \cdot (P_{INT} + \alpha \cdot r_{p2w} \cdot C_{wire} \cdot {V_{DD}}^2 \cdot f_{clk}) + \Delta_{wire} \cdot \alpha \cdot C_{wire} \cdot {V_{DD}}^2 \cdot f_{clk}, \quad (3.2)$$

where Δ_{cell} denotes the standard cell area saving from M3D ICs over the 2D counterparts, and Δ_{wire} denotes the wire-length saving in the M3D IC. This simple linear model gives useful insight in explaining the power saving trends across technology nodes and frequen-

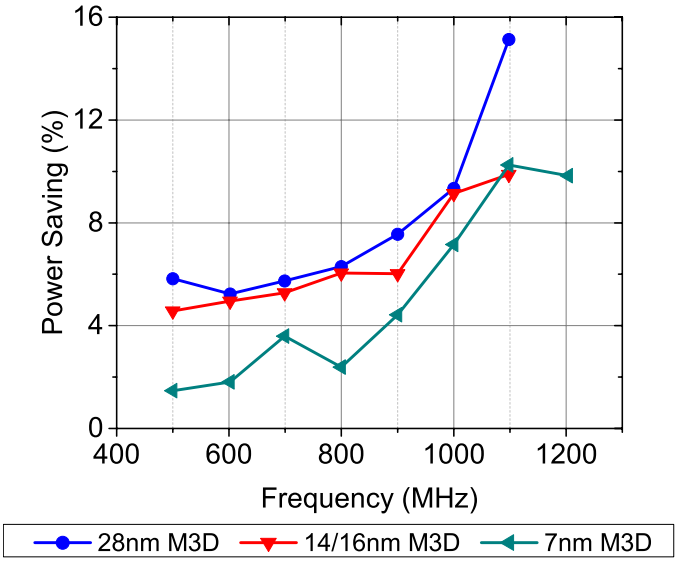

Figure 3.3: Power saving of M3D ICs over 2D ICs in 28nm, 14/16nm and 7nm technology nodes

cies.

Analysis of Trends

As can be seen from Figure 3.4, at a given frequency, the wire-length saving (Δ_{wire}) as well as the standard cell area saving (Δ_{cell}) is nearly the same across all the three technology nodes.

As the clock frequency is swept, the wire-length saving (Δ_{wire}) does not vary by a large magnitude, ranging between 20% to 25% as shown in Figure 3.4. However, with increasing clock frequency, 2D ICs utilize more buffers and higher drive-strength cells to meet timing whereas M3D ICs can meet timing with lesser number of buffers and lower drive-strength cells because of the wire-length saving. Hence, the standard cell area saving (Δ_{cell}) increases from 2% up to 10$\sim$12% with increasing frequency. With these observations, Equation (3.2) is modified to denote Δ_{cell} as a function of f_{clk} in order to reflect the impact of frequency on standard cell area savings.

$$\begin{aligned}\Delta P_{dyn} =& \Delta_{cell}(f_{clk}) \cdot (P_{INT} + \alpha \cdot r_{p2w} \cdot C_{wire} \cdot {V_{DD}}^2 \cdot f_{clk}) \\ &+ \Delta_{wire} \cdot \alpha \cdot C_{wire} \cdot {V_{DD}}^2 \cdot f_{clk}\end{aligned} \tag{3.3}$$

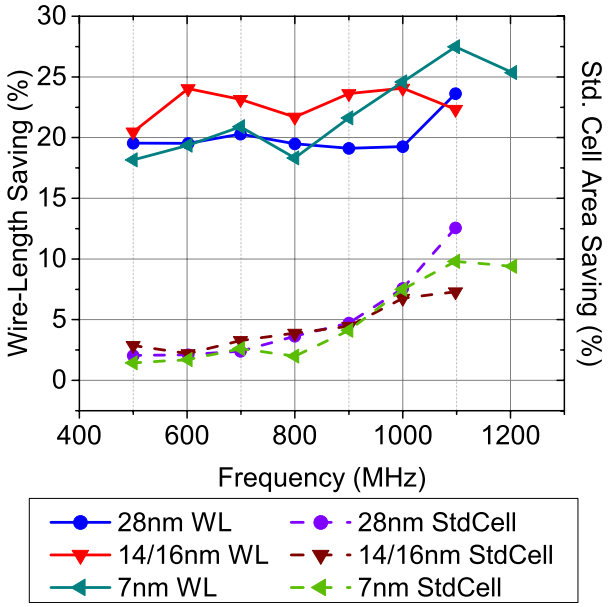

Figure 3.4: Impact of M3D ICs on the wire-length (solid lines) and standard cell area (dotted lines) savings over 2D ICs in $28nm$, $14/16nm$ and $7nm$ technology nodes

M3D Power Saving at Low Frequency At low frequencies ($500MHz$), Δ_{cell} is small (= 2.5%) while Δ_{wire} is much higher. Hence, most of the power saving in M3D ICs comes from reduction in wire capacitance switching power. Figure 3.5 (a) shows the normalized power components of the 2D and M3D ICs across technology nodes at the minimum ($500MHz$) and maximum ($1.1/1.2GHz$) frequencies. This figure clearly shows that internal power is the dominant portion of the total power accounting for nearly 50% of the total across all frequencies and technology nodes. The rest is split between pin capacitance switching and wire capacitance switching power with leakage power taking up the smallest portion. It is important to note that the pin capacitance switching power and internal power are both related to the number and size of gates used in the design. Power saving due to reduction in wire capacitance switching power is determined by r_{p2w} in the design. Hence, even with 20$\sim$25% wire length reduction, the total power saving at low frequencies ranges between 1.5% for the $7nm$ technology node to 6% for the $28nm$ technology node because wire capacitance switching power is a small portion of the total power. The $28nm$ M3D IC has better power saving at $500MHz$ because it has a larger wire capacitance to

pin capacitance ratio as shown in Figure 3.6.

This difference in pin capacitance versus wire capacitance from $28nm$ to 14/16nm node can be attributed to the difference in gate capacitance associated with planar MOSFETs and 3D FinFETs. FinFET based technologies have higher gate capacitance due to the 3D fin structure and the introduction of local interconnect MOL layers that contact the device terminals to M1. This observation that planar MOSFET based designs are more likely benefit from M3D ICs compared to FinFET based designs at advanced nodes, is a key finding of this work.

Another point to note is that with process scaling wire RC parasitic, especially resistance, increases per unit length. Improving drive-strength of transistors at advanced nodes like $7nm$ is extremely challenging. As the ratio of transistor drive versus wire-load decreases at scaled nodes, implementation tools end up using larger cells to drive the same wire-length, hence, effectively increasing r_{p2w}. Hence, technologies with larger transistor fanouts will benefit more from M3D ICs.

M3D Power Saving at High Frequency At high operating frequencies, as Δ_{cell} increases, it affects both pin capacitance switching power and internal power. As evident from Figure 3.5, internal power and pin capacitance switching power can contribute up to 70% of the total power of the 2D IC at high frequencies. Hence, the total power savings at maximum frequencies approach 10% or more as M3D ICs benefit from reduction in all power components, predominantly internal power and pin capacitance switching power.

In order to understand the impact of frequency on M3D power savings, a hypothetical scenario is considered when, with increasing clock frequency, $\Delta_{cell}(f_{clk}) = \Delta_{wire}$. At this frequency point, Equation (3.3) can be modified to the following expression.

$$\Delta P_{dyn} = \Delta_{wire} \cdot (P_{INT} + \alpha \cdot C_{tot} \cdot {V_{DD}}^2 \cdot f_{clk}), \tag{3.4}$$

where C_{tot} is total capacitance ($C_{pin} + C_{wire}$) of a 2D IC. At this clock frequency, the M3D power saving does not depend on r_{p2w}. Moreover, as discussed previously, P_{INT} being the

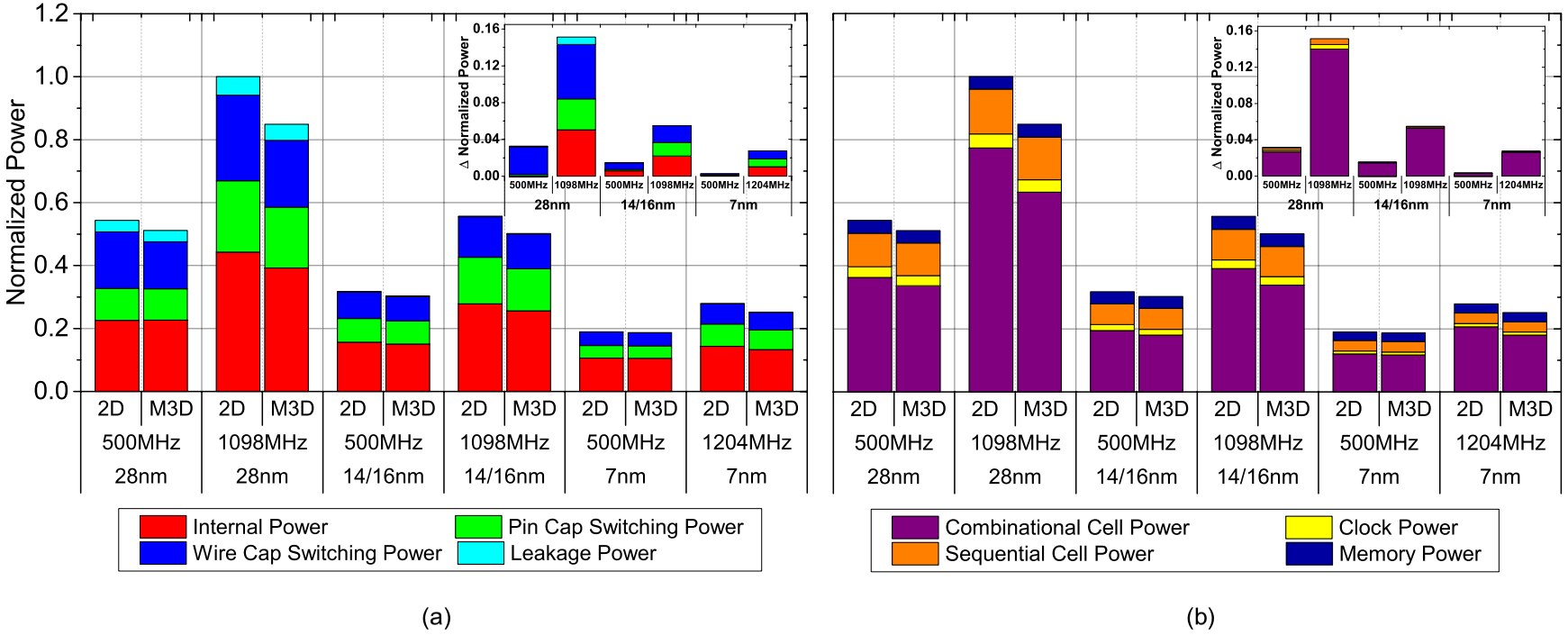

Figure 3.5: Power breakdown into (a) internal power, pin capacitance switching power, wire capacitance switching power, and leakage power, (b) combinational cell power, clock power, sequential cell power, and memory power at the minimum and maximum frequencies of each technology nodes. The inset plots shows the power reduction of M3D ICs for each power component.

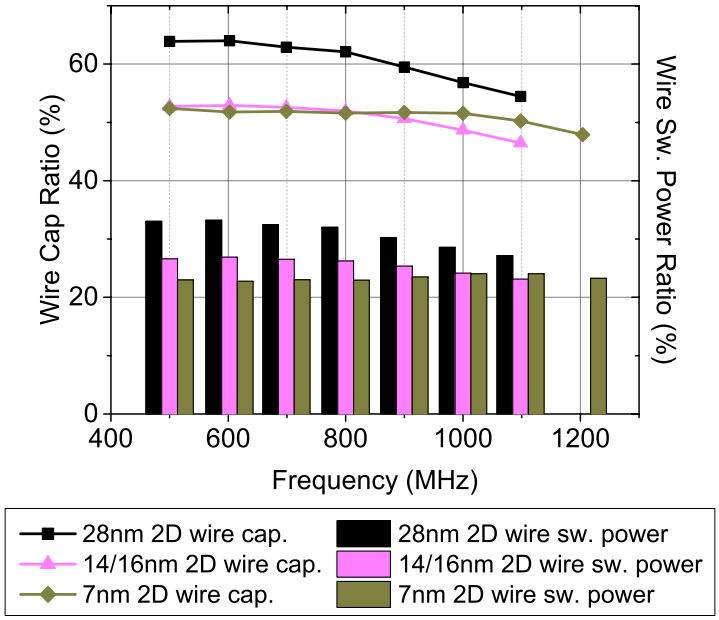

Figure 3.6: Wire capacitance to total capacitance ratio, and net switching power to total power ratio in 2D ICs in $28nm$, 14/16nm and $7nm$ technology nodes

dominant component of the total power, M3D power saving depends more on Δ_{cell} and the ratio of internal power versus net switching power than Δ_{wire} or r_{p2w}.

Figure 3.5 (b) shows power breakdown according to the type of cells. As the number of hard macros (e.g., memory blocks) and sequential cells are fixed, the power consumed by these cells do not change in M3D ICs. On the other hand, power consumed by combinational cells and clock signal can be reduced effectively in M3D ICs utilizing lower number of buffers and using lower drive-strength cells.

Table 3.2 shows all the important design metrics of both 2D and M3D ICs across foundry $28nm$, 14/16nm, and the predictive $7nm$ technology nodes at $1.1GHz$. Since $1.1GHz$ is the maximum frequency for $28nm$ and 14/16nm implementation, and the second highest for $7nm$ design, the significant standard cell area saving as well as wire-length saving is achieved with M3D ICs.

Since the operating clock frequency is high, M3D ICs save the standard cell area by 9.9% on average for the three implementations, resulting in the internal power and pin capacitance switching power savings. Although the ratios of internal power and pin capac-

Table 3.2: Normalized iso-performance design and power metric comparison of 2D and M3D ICs with application processor in $28nm$, 14/16nm, and $7nm$ technology nodes. All values are normalized to corresponding $28nm$ 2D parameters. Capacitance and power values are normalized to $28nm$ 2D total capacitance and $28nm$ 2D total power, respectively.

	normalized 2D IC			M3D IC percentage change from 2D		
parameters	$28nm$	14/16nm	$7nm$	$28nm$	14/16nm	$7nm$
footprint	1x1	0.64x0.64	0.41x0.35	-51.1%	-50%	-54.7%
density	1	0.899	0.803	-10.9%	-8.9%	-12.3%
cell count	1	1.029	1.251	-7.8%	-7.3%	-9.5%
std. cell area	1	0.32	0.085	-12.6%	-7.3%	-9.8%
wire-length	1	0.649	0.437	-23.6%	-22.3%	-27.5%
wire cap	0.544	0.328	0.207	-23.3%	-13.1%	-13.2%
pin cap	0.456	0.378	0.205	-16.5%	-9.1%	-12%
total cap	**1**	**0.706**	**0.412**	**-20.2%**	**-11%**	**-12.6%**
internal power	0.443	0.278	0.136	-11.4%	-7.9%	-8.6%
wire cap switching power	0.271	0.129	0.063	-21.8%	-14%	-12.7%
pin cap switching power	0.227	0.148	0.062	-14.9%	-10.1%	-11.5%
leakage power	0.059	0.001	0.001	-13.4%	-5%	-3.2%
total power	**1**	**0.557**	**0.262**	**-15.1%**	**-9.9%**	**-10.3%**

itance switching power on 2D ICs versus M3D ICs (11.4% and 14.9% in the $28nm$ designs) is smaller than wire capacitance switching power ratio (= 21.8%), since those components account for more than 70% of the total power, they have a bigger contribution to the total power savings.

3.3 A Design-Aware Partitioning Approach to Monolithic 3D IC with 2D Commercial Tools

Based on the observations in the previous section, a new methodology called '**cascade-2D design flow**' to implement M3D ICs using 2D commercial tools is presented in this work. **Cascade-2D design flow** utilizes a design-aware partitioning scheme where functional modules with very large number of connections are partitioned into separate tiers. One of the main advantages of this flow is that it is extremely flexible and is partition-scheme agnostic, making it an ideal methodology to evaluate different tier partitioning algorithms.

The MIVs are modeled as sets of anchor cells and dummy wires, which enable to implement and optimize both top and bottom tiers simultaneously in a 2D IC. **Cascade-2D design flow** reduces standard cell area effectively, resulting in significantly better power savings than **shrunk-2D design flow**. Experimental results show that M3D ICs implemented with **cascade-2D design flow** (i.e., **cascade-2D** M3D ICs) can achieve up to 4× better power savings compared to those with **shrunk-2D design flow** (i.e., **shrunk-2D** M3D ICs), while using an order of magnitude less MIVs. In the best case scenario, **cascade-2D** M3D ICs result in 25% higher performance at iso-power and up to 20% power reduction at iso-performance compared to 2D ICs. Additionally, by leveraging smaller standard cells, M3D ICs can save up to 10% die area which directly translates to reduced costs.

Figure 3.7 shows the 'cut-and-slide' methodology of **cascade-2D design flow** with sets of anchor cells and dummy wires. As can be clearly seen, the anchor cells and dummy wires model MIVs, and the *cascade-2D design* implementation in Figure 3.7 (a) is functionally

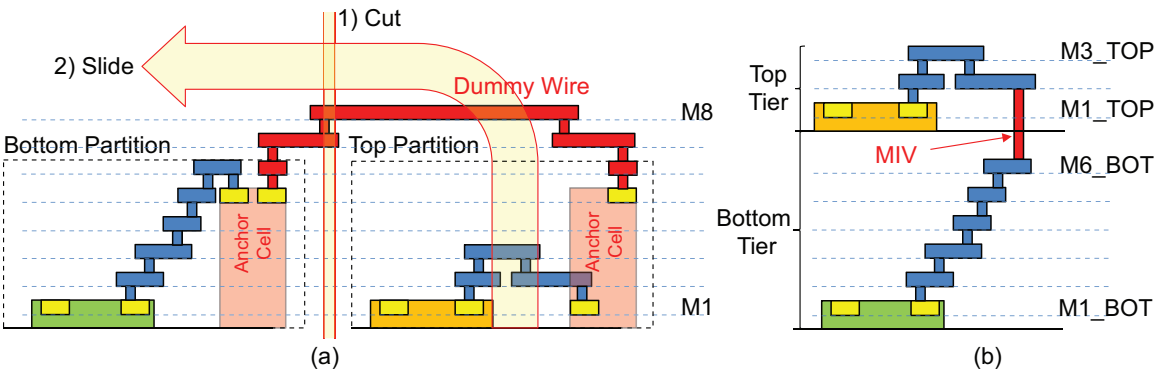

Figure 3.7: M3D IC design scheme of **cascade-2D design flow**. (a) A *cascade-2D design* implementation with a set of anchor cells and dummy wires, which models MIVs, and (b) the equivalent M3D IC.

Table 3.3: Qualitative comparison of **cascade-2D design flow** and **shrunk-2D design flow**

cascade-2D design flow	**shrunk-2D design flow**
• Support block- and gate-level M3D ICs • Capable of handling RTL-level constraints • Highly flexible; can implement any partitioning algorithm • Designer has complete control over tier-assignment of cells/blocks • Implements top and bottom tier in a single design • Buffer insertion based on actual technology parameters	• Support only gate-level M3D ICs • Cannot handle RTL-level constraints • Implements area-balanced min-cut algorithm for partitioning cells • Designer controls bin-size but not actual tier-assignment of gates • Implements top and bottom tier separately • Buffer insertion based on shrunk technology parameters

equivalent to the M3D IC in Figure 3.7 (b).

3.3.1 Implementation Methodology

Table 3.3 presents a qualitative comparison of **cascade-2D design flow** with **shrunk-2D design flow**. Figure 3.8 shows the flow diagram of this methodology. First, functional blocks are partitioned into two groups, the top and bottom group, creating signals crossing the two groups, which become MIVs in an M3D IC. Then, the location of MIVs are determined, and lastly, a *cascade-2D design* is implemented with sets of anchor cells and dummy wires in 2D space, which is equivalent to the final M3D IC.

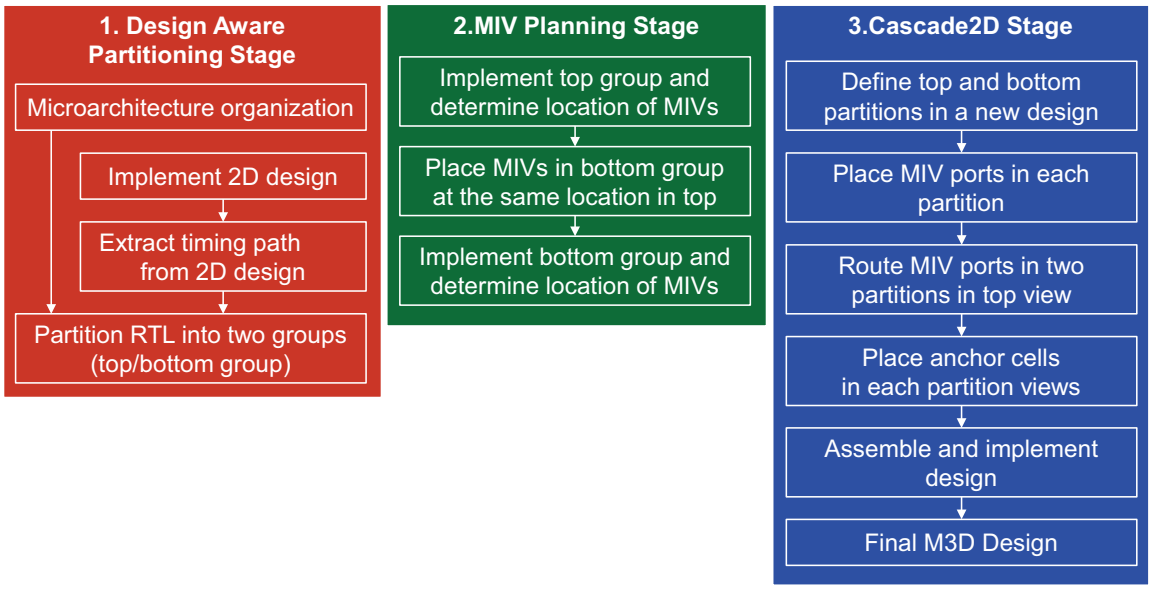

Figure 3.8: Flow diagram of **cascade-2D design flow**

Design-Aware Partitioning Stage

In this step, RTL is partitioned into two groups, the top and bottom group, which represent the top and bottom tier of an M3D IC, respectively. The partition can be performed in two ways: (1) based on the organization of the design micro-architecture and (2) by extracting design information from the 2D implementation.

Because M3D ICs offer vertical integration of cells, power and performance improvement is achieved by placing inter-communicating functional modules separated by a large distance in the xy-plane in a 2D IC on separate tiers and reducing the distance by utilizing the z-axis in an M3D IC. With a detailed understanding of the micro-architecture organization, functional modules can be pre-partitioned into separate tiers. For example, consider two functional modules whose connecting signals have a tight-timing budget (e.g., a data path unit and its register bank). Placing these modules into separate tiers and connecting them with MIVs can help reduce the wire-length.

In case it is non-trivial to partition based on the understanding of micro-architectural organization, the design information from a 2D implementation can be utilized to help guide

the partitioning process. By extracting timing paths from a 2D IC, the number of timing paths crossing each pair of functional modules can be quantified, which is called 'degree of connectivity' between functional modules. The standard cell area of each functional module is also extracted from the 2D IC for cell area balancing between the tiers.

After obtaining the degree of connectivity of functional modules and their cell area, the design is partitioned into two groups based on the following criteria:

- Balance the cell area of the top and bottom group
- Maximize the number of timing paths crossing the two groups

These criteria helps (1) the functional blocks, which have a very high degree of connectivity, be placed into separate tiers and minimize the distance between them, and (2) balance the standard cell area of the two tiers. Figure 3.9 shows an example of design-aware partitioning. Module A and B are fixed on two different groups based on organization of the design micro-architecture, module C, D, E, and F are partitioned maximizing the number of timing paths crossing two groups and balancing cell area of two groups. It should be emphasized, however, that **cascade-2D design flow** is extremely flexible, and can incorporate any number of constraints for partitioning cells or modules into separate tiers. Depending on the type of design, the designer may wish to employ different partitioning criteria than presented here and the subsequent steps (MIV Planning Stage and Cascade-2D Stage) would remain the same. Hence, this flow is an ideal platform to evaluate different tier partitioning schemes for M3D ICs.

At this stage, it is important to understand that there are two types of IO ports in the design. There are a set of IO ports that were created because of the 'design-aware partitioning' step. These IO ports connect the top and bottom groups of the design, and they are referred as MIV ports in rest of the work since they eventually become MIVs in an M3D IC. Additionally, there exist a set of IO ports for the top-level pre-partitioned design. These are same as the conventional IO ports of 2D ICs.

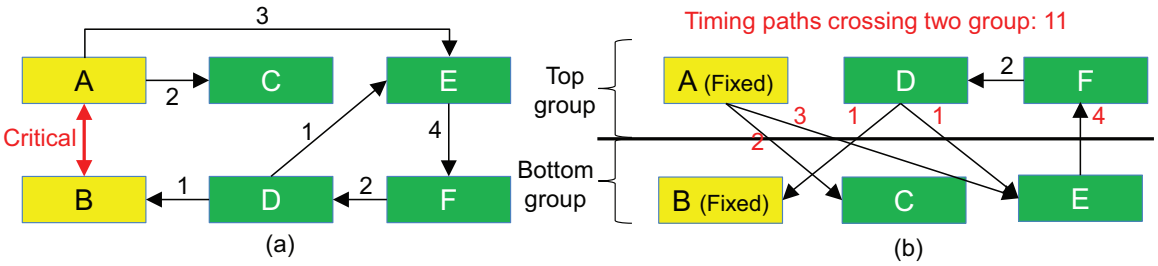

Figure 3.9: An example of the design-aware partitioning scheme of **cascade-2D design flow** (a) Pre-partitioned modules (yellow box), and degree of connectivity (numbers on the arrows) of rest of modules (green box). (b) Result of the design-aware partitioning

MIV Planning Stage

After partitioning the RTL into the top and bottom groups, the location of MIVs are determined. First, the top group is implemented, and MIVs ports are placed above their driving or receiving cells on the top routing metal layer, so that wire-length between MIV ports and relevant cells are minimized. As explained previously, MIV ports are actually IO ports that connect the top and bottom groups. Leverage the fact that all cell placement algorithms in commercial EDA tools tend to place cells close to IO ports to minimize timing, the bottom group is implemented using the location of MIVs determined from the top group implementation. In this way, the cell placement of the top group guides the cell placement of the bottom group using the pre-fixed MIV ports.

In this work, the IO ports of the top-level design are assumed to be connected only to the top tier in M3D ICs. Therefore, it is possible that some IO signals need to be directly connected to functional modules in the bottom group. These feed-through signals will not have any driving or receiving cells on the top group. Hence, the MIV ports for those signals cannot be placed with top group implementation and are determined during the bottom group implementation.

Figure 3.10 shows the location of MIVs after implementing the bottom group. After obtaining the location of complete set of MIVs, standard cell placement in the top and bottom group implementation is discarded, and only the MIV locations are retained.

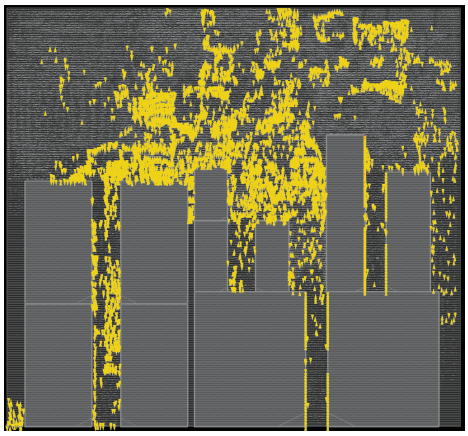

Figure 3.10: Location of MIVs (yellow dots) after completing MIV planning stage in **cascade-2D design flow**

Cascade-2D Stage

In this step, a *cascade-2D designs* is implemented, which models an M3D IC in a single 2D design with sets of anchor cells and dummy wires, using partitioning technique supported in Cadence® Innovus™.

First, a new die with both tiers placed side-by-side is created, with the same total area as the original 2D design. Top and bottom partitions is defined in the die, and a hard fence for placement is set, so that cells in the top partition are placed only on the top half of the die, and cells in the bottom partition only on the bottom half of the die. Then two hierarchies of the design are created as follows:

- 1st level of hierarchy: Top view, which contains only two cells, top-partition cell and bottom-partition cell. These two cells contain pins which represent MIVs for the top and bottom tier, respectively.
- 2nd level of hierarchy: Top partition-cell, which contains the top partition view where standard cells from the top group are placed and routed
- 2nd level of hierarchy: Bottom partition-cell, which contains the bottom partition view where standard cells from the bottom group are placed and routed

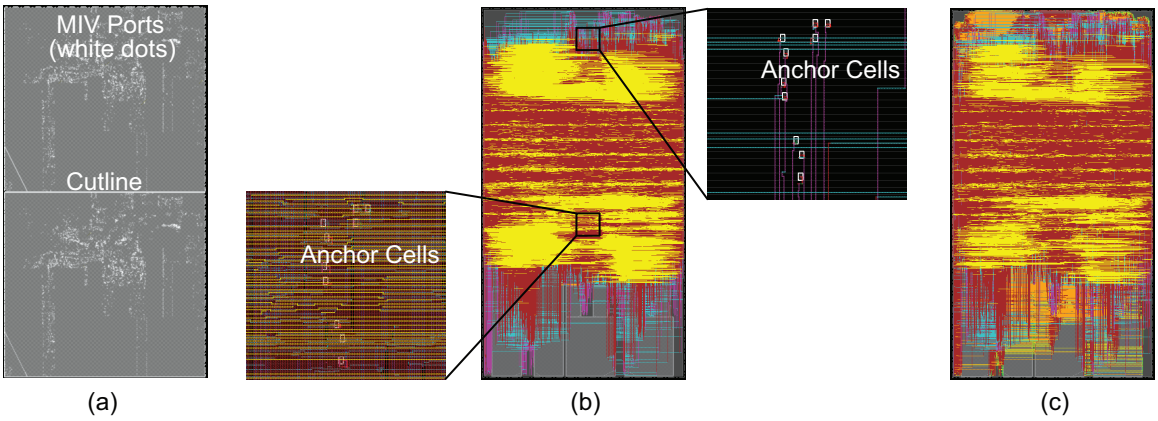

Figure 3.11: GDS layouts in each steps in **cascade-2D design flow**. (a) top view after placing pins for MIVs, (b) after assembling top view and top and bottom partition view, (c) after implementing *cascade-2D designs*

In the top view, pins are placed, representing MIVs, in the top-partition cell and bottom-partition cell on the top routing metal layer (i.e., M6 in Figure 3.7). The pin locations are the same as the MIV location derived in MIV Planning Stage. Figure 3.11 (a) shows placed pins for MIVs in the top view.

Then, using 3∼4 additional metal layers above the top routing metal layer used in the actual design, (i.e., M7∼M8 in Figure 3.7), the pins on the top-partition cell and bottom-partition cell are routed and connected. As the location of the pins are identical in the x-axis in the top and bottom-partition cells, the routing tool creates long vertical wires crossing two partition cells. These additional 3∼4 metal layers used to connect the pins of the top and bottom partitioning cells are called 'dummy wires' because their only function is to get logical connection between the two tiers in the physical design. The delay and RC parasitics associated with these wires will not be considered in the final M3D IC.

In an M3D IC, the last metal layer of the bottom tier is connected to the first metal layer of the top tier using an MIV. To emulate this connectivity in a 2D IC where the top and bottom tier are placed adjacent to each other, a mechanism to connect M1 in the top partition view with M6 in the bottom partition view is required. This is achieved through, 'anchor cells'. An anchor cell is a dummy cell which implements buffer logic. Anchor cells

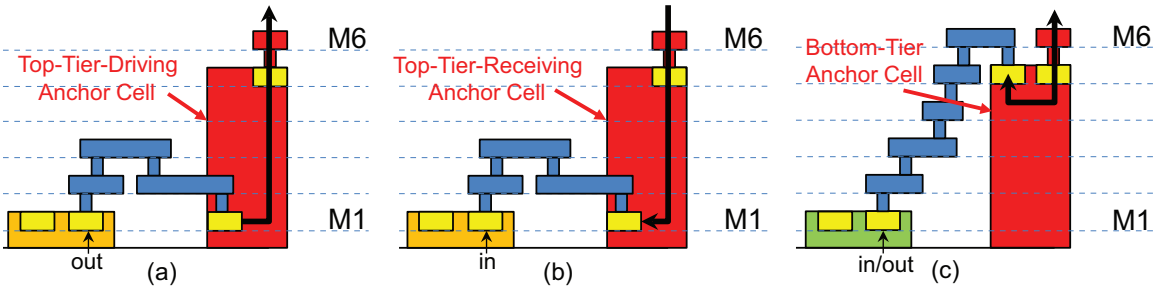

Figure 3.12: Three types of anchor cells: (a) a top-tier-driving anchor cell (b) a top-tier-receiving anchor cell, and (c) a bottom-tier anchor cell

model zero-delay virtual connection between a dummy wire and one of the metal layers. After connecting the two partition cells with dummy wires, anchor cells are placed below the pins in each partition view. In this step, only anchor cells are placed but not logic cells.

Depending on the partition using anchor cells and metal layer to which a dummy wire needs to be virtually connected, three flavors of anchor cells exist: (1) top-tier-driving anchor cells (Figure 3.12 (a)), which are placed in the top partition, receiving signals from M1 of top partition, and driving a dummy wires, (2) top-tier-receiving anchor cells (Figure 3.12 (b)), which sends signal in the reverse direction, and (3) bottom-tier anchor cells (Figure 3.12 (c)), which are placed in the bottom partition, connecting a dummy wire to top metal layer of the bottom partition. After placement, anchor cells and the corresponding MIV ports are connected.

Next all hierarchies are flattened, i.e., the top view and both partition views are assembled projecting all anchor cells in two partition views and dummy wires in the top view into a single design. Figure 3.11 (b) shows the assembled design.

With the assembled design, the delay of dummy wires is set to zero, and anchor cells and dummy wires are set to be fixed, so that their location cannot be modified. These sets of anchor cells and dummy wires effectively act as ‘wormholes’ which connect M1 of the top partition and top routing metal layer of the bottom partition without delay emulating the behavior of MIVs (the MIV RC parasitics are added in the final timing stage).

Then regular 2D IC design flow is performed, which involves all the design stages

including placement, post-placement optimization, CTS, post-CTS optimization, routing, and post-route optimization. Owing to (1) 'wormholes', which provide virtual connection between M1 of the top partition and top routing metal layer of the bottom partition, and (2) the hard fence, which sets the boundary for top and bottom partition, the tool places each tier in its separate 2D partitioned space with virtual connections between them, resulting in a *cascade-2D design*.

CTS in **cascade-2D design flow** is performed as regular 2D IC design flow. A clock signal is first divided into two branches in the top partition. One of the branches is used for generating the clock tree in the top partition, and the other branch is connected to the bottom partition through a set of anchor cells and a dummy wire, and used for generating the clock tree in the bottom partition.

Figure 3.11 (c) shows the resulting *cascade-2D design*. Although the delay of dummy wires is set to zero their RC parasitics still exist in this stage of the design. Therefore, the *cascade-2D design* is again split into top and bottom partitions, pushing all cells and wires to the corresponding partitions except dummy wires. Then, RC parasitics for each partition are extracted. The final M3D IC is created by connecting these two extracted designs with MIV RC parasitics. Timing and power analysis is done on the final M3D IC.

3.3.2 Impact of New Monolithic 3D IC Design Flow

Same experimental setup as described in Section 3.2.1 is used for gauge the benefits of **cascade-2D design flow** over **shrunk-2D design flow**.

Power and Performance Benefit

Figure 3.13 shows the GDS layouts of 2D and **cascade-2D** M3D ICs of the commercial, in-order, 32-bit application processor on target frequency of 1.0GHz in 28nm, 14/16nm as well as 7nm technology nodes.

Timing analysis of the 2D IC indicates functional module A and B in Figure 3.14 (a)

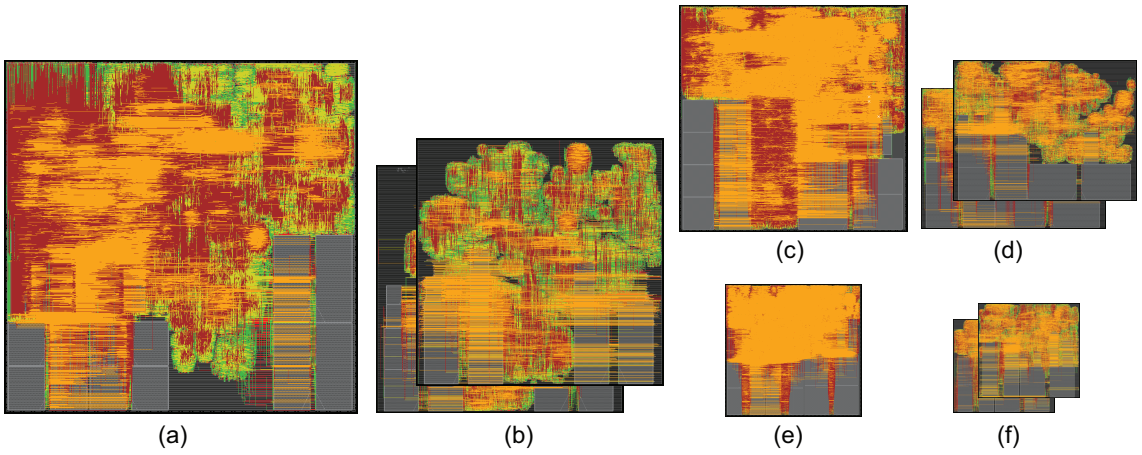

Figure 3.13: GDS layouts of (a) 28nm 2D, (b) 28nm **cascade-2D** M3D, (c) 14/16nm 2D, (d) 14/16nm **cascade-2D** M3D, (e) 7nm 2D and (f) 7nm **cascade-2D** M3D ICs of the application processor at 1.0GHz

have a large number of timing paths crossing them. In the **cascade-2D** M3D IC, those modules are floorplanned on top of each other minimizing the distance between them using MIVs, whereas those functional modules are floorplanned side-by-side in the 2D IC. This vertical integration reduces the wire-length of signals crossing the modules as well as the standard cell area of the modules because of reduced wire RC parasitics.

The normalized total power consumption of the 2D and **cascade-2D** M3D ICs across technology nodes are shown in Figure 3.15. **Cascade-2D** M3D ICs consume less power in all cases. Hence, at iso-power, M3D ICs run at higher frequencies compared to the 2D ICs. For example, in the 14/16nm technology node, M3D ICs can have 25$\%$ higher performance at the same total power compared to the 2D ICs. Figure 3.16 shows power saving comparison between **cascade-2D** M3D and **shrunk-2D** M3D ICs from their 2D counterparts. **Cascade-2D** M3D ICs show up to 3$\sim$4$\times$ better power saving than **shrunk-2D** M3D ICs depending on the technology node and design frequency. In the best case scenario, M3D IC shows 20$\%$ power reduction than the 2D IC (14/16nm technology node at 1.1GHz frequency) at the same performance point.

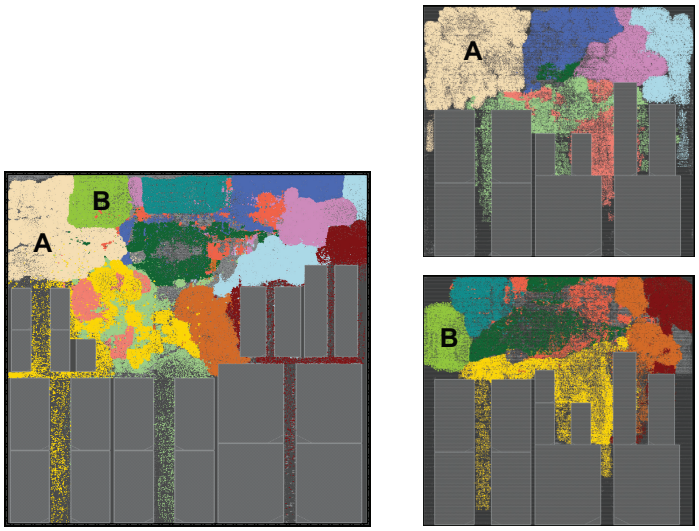

Figure 3.14: Color map of functional modules in application processor $7nm$ (a) 2D IC and (b) **cascade-2D** M3D IC of commercial application processor at $1.0GHz$

Table 3.4: Number of MIVs in application processor M3D ICs in $28nm$, 14/$16nm$, and $7nm$ technology nodes

MIV count	$28nm$	14/$16nm$	$7nm$
cascade-2D M3D IC	7,545	7,545	7,545
shrunk-2D M3D IC	164,553	120,770	99,587

Comparison to Shrunk-2D Design Flow

As discussed in Section 2.3.2, the primary advantage of **shrunk-2D** M3D ICs comes from reduced wire-length, which results in reduced wire capacitance switching power dissipation. As shown in Figure 3.17, **shrunk-2D** M3D ICs reduce wire-length by 20$\sim$25% consistently across technology nodes and frequencies. Wire-length reduction is mainly attributed to vertical integration between cells through MIVs. Table 3.4 compares the number of MIVs **shrunk-2D** M3D and **cascade-2D** M3D ICs. Since **shrunk-2D design flow** partitions cells into two tiers whereas **cascade-2D design flow** partitions functional blocks, the number of MIVs in **shrunk-2D** M3D ICs is an order of magnitude higher than that in **cascade-2D** M3D ICs. Better wire-length savings using **shrunk-2D design flow** can be attributed to the large number of MIVs.

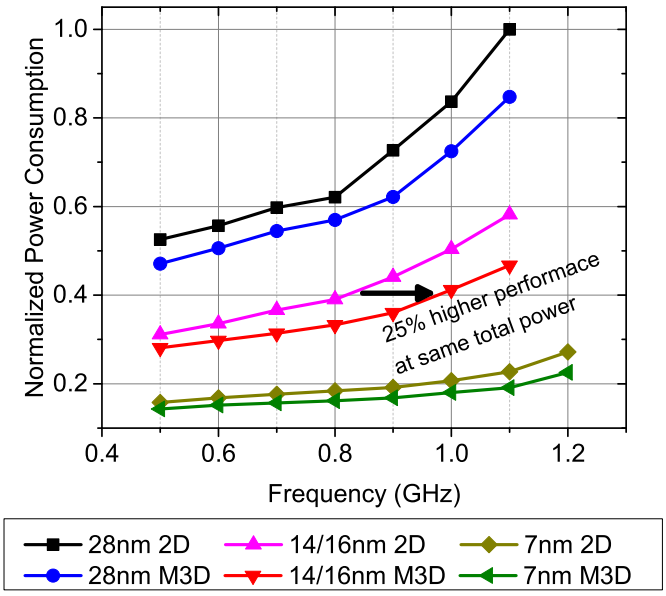

Figure 3.15: Normalized power consumption of 2D and **cascade-2D** M3D ICs in 28nm, 14/16nm, and 7nm technology nodes

The large number of MIVs in **shrunk-2D** M3D ICs helps to reduce wire-length, but it also increases the total capacitance of MIVs, limiting the wire capacitance reduction. As shown in Table 3.5, although **shrunk-2D** M3D ICs reduce more wire-length than **cascade-2D** M3D ICs in 14/16nm and 7nm designs, the wire capacitance reduction of **cascade-2D** M3D ICs higher than **shrunk-2D** M3D ICs. Additionally, there is a negative impact of the large number of MIVs on the wire capacitance mainly because of the bin-based partitioning scheme of **shrunk-2D design flow** [5]. While the bin-based partitioning helps distribute cells evenly on both tiers, it has a tendency to partition cells connected using local wires into two tiers, increasing the wire capacitance.

On the other hand, **cascade-2D** M3D ICs save their power mainly by reducing standard cell area. **Shrunk-2D design flow** uses a *shrunk-2D design* to estimate the wire-length and the wire RC parasitics of the resulting M3D IC. However, while shrinking process geometries, minimum width of each metal layer is also scaled, and extrapolation is performed by tools during RC extraction of wires. This extrapolation tends to overestimate wire RC parasitics, especially in scaled technology nodes, which results in a large number of buffers

Table 3.5: Normalized iso-performance comparison of 2D, **shrunk-2D** M3D and **cascade-2D** M3D ICs with application processor in $28nm$, 14/16nm, and $7nm$ technology nodes. All values are normalized to corresponding $28nm$ 2D parameters. Capacitance and power values are normalized to $28nm$ 2D total capacitance and $28nm$ 2D total power, respectively.

	normalized 2D			**shrunk-2D** percentage change from 2D			**cascade-2D** percentage change from 2D		
parameters	$28nm$	14/16nm	$7nm$	$28nm$	14/16nm	$7nm$	$28nm$	14/16nm	$7nm$
std. cell area	1	0.331	0.077	-7.6%	-6.8%	-7.5%	-9.5%	-11.9%	-8.8%
wire-length	1	0.728	0.404	-19.3%	-24.1%	-24.6%	-11.9%	-22.6%	-12.2%
wire cap	0.531	0.375	0.205	-18.1%	-14.2%	-13.7%	-9.5%	-19.7%	-19.2%
pin cap	0.469	0.422	0.203	-12.1%	-6.3%	-9.7%	-11.1%	-13.2%	-7.9%
total cap	**1**	**0.797**	**0.408**	**-15.5%**	**-10.1%**	**-11.7%**	**-9.6%**	**-15.2%**	**-12.9%**
internal power	0.428	0.282	0.128	-4.8%	-7.6%	-4.7%	-14.5%	-15.2%	-11.1%
net switching power	0.505	0.318	0.119	-13.4%	-10.6%	-10.1%	-13.0%	-20.8%	-15.1%
leakage power	0.066	0.002	0.000	-7.7%	-4.0%	-2.0%	-9.5%	-7.7%	-2.8%
total power	**1**	**0.602**	**0.247**	**-9.3%**	**-9.1%**	**-7.2%**	**-13.4%**	**-18.1%**	**-13%**

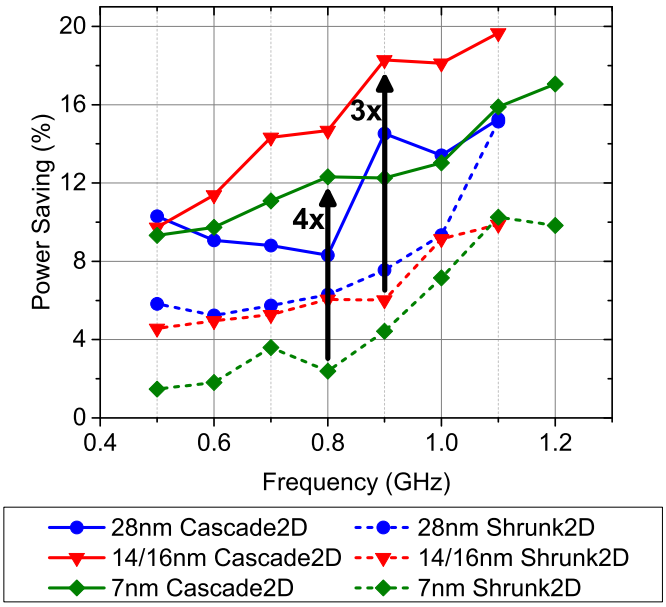

Figure 3.16: Power saving of **cascade-2D** M3D (solid lines) and **shrunk-2D** M3D (dotted lines) ICs over 2D ICs in 28nm, 14/16nm, and 7nm technology nodes

inserted in a design to meet timing. In **cascade-2D design flow**, buffers are inserted while implementing and optimizing top and bottom partitions simultaneously with actual process geometries, **cascade-2D design flow** achieves more standard cell area than **shrunk-2D design flow** as shown in Figure 3.18.

With a reduction in standard cell area, the cell density of the M3D IC reduces as well. Hence, leveraging this feature of M3D ICs to increase cell density and reduce die area, two separate M3D ICs are implemented using **cascade-2D design flow**, one with the same total die area as the 2D IC and another with 10% reduced area. Table 3.6 shows that similar power savings can be achieved with a reduced die area M3D IC. The ability to get reduced die area makes M3D stacking technology extremely attractive for main-stream adoption because less area directly translates to reduced costs.

As shown in Equation (3.3), standard cell area reduction affects both internal power, pin capacitance switching power reduction, whereas wire-length reduction reduces only wire capacitance switching power. Figure 3.19 shows the power breakdown of 2D, **cascade-2D** M3D, and **shrunk-2D** M3D ICs. As shown in the figure, the internal power and pin

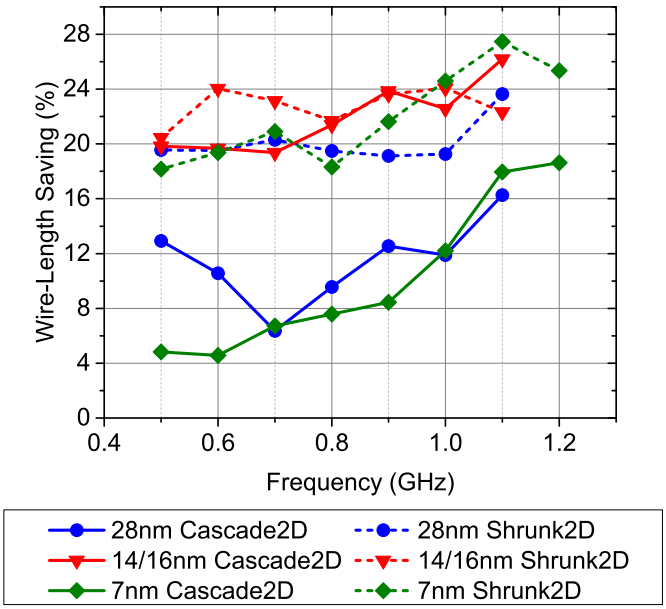

Figure 3.17: Wire-length reduction comparison between **cascade-2D** (solid lines) and **shrunk-2D** (dotted lines) M3D ICs over 2D ICs

Table 3.6: Normalized iso-performance power metric comparison of 2D and **cascade-2D** M3D IC with the same and 0.9× footprint at 1.1GHz in predictive 7nm technology node

parameters	2D	**cascade-2D** M3D	
die area	1	1	0.9
density	69.7%	63.2%	71.1%
total power	1	0.841	0.871

capacitance switching power, which depends on the standard cell area, account for over 70% of the total power, and they contribute even more in 14/16nm and 7nm designs. **Cascade-2D** M3D ICs reduce more standard cell area compared to **shrunk-2D** M3D ICs by attacking 70% of the total power; they achieve better power savings consistently, even though the wire-length reduction of **cascade-2D** M3D ICs is less than **shrunk-2D** M3D ICs.

Table 3.7 shows the comparison of run-time between **cascade-2D design flow** and **shrunk-2D design flow**. For **shrunk-2D design flow**, the design library with shrunk geometry is assumed to be available. The total run-time for each flow is comparable. It is important to note that both flows need a reference 2D IC. The 2D IC is needed in **shrunk-2D design flow** to evaluate the quality of the final M3D IC, while it is useful in **cascade-2D**

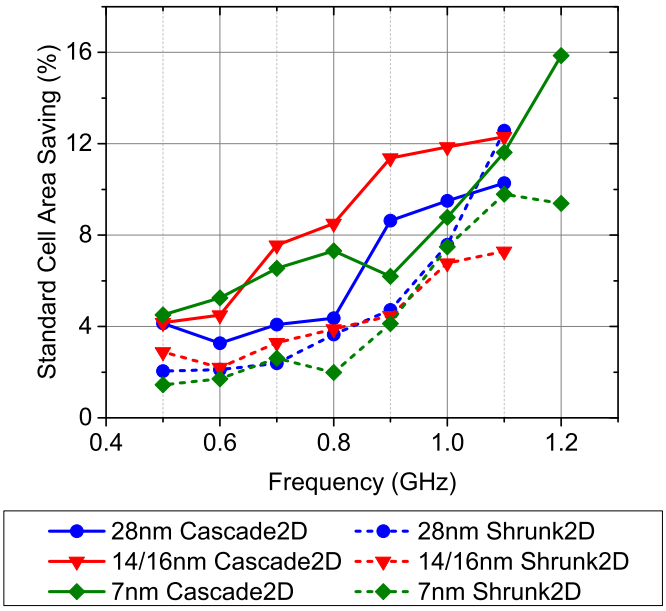

Figure 3.18: Standard cell area saving in **cascade-2D** (solid lines) and **shrunk-2D** (dotted lines) M3D ICs over 2D ICs

Table 3.7: Run-time comparison between **shrunk-2D** and **cascade-2D design flow** with application processor in $7nm$ technology node

shrunk-2D design flow		**cascade-2D design flow**	
step	run-time	step	run-time
1. shrunk-2D impl.	5hr	1. design-aware part.	0.5hr
2. gate-level part.	0.5hr	2. MIV plan	2hr
3. MIV plan	0.5hr	3. cascade-2D impl.	4.5hr
4. top/bottom tier impl.	1.5hr	-	-
total	7.5hr	total	7hr

design flow to extract timing and standard cell area information for the design-aware partitioning step.

3.4 Summary

3.4.1 Benefit Trends of Monolithic 3D ICs across Technology Nodes

The observations on the power benefit trends of M3D ICs across technology nodes are summarized as follows:

- Designs using planar MOSFETs are more likely to gain from M3D stacking technology

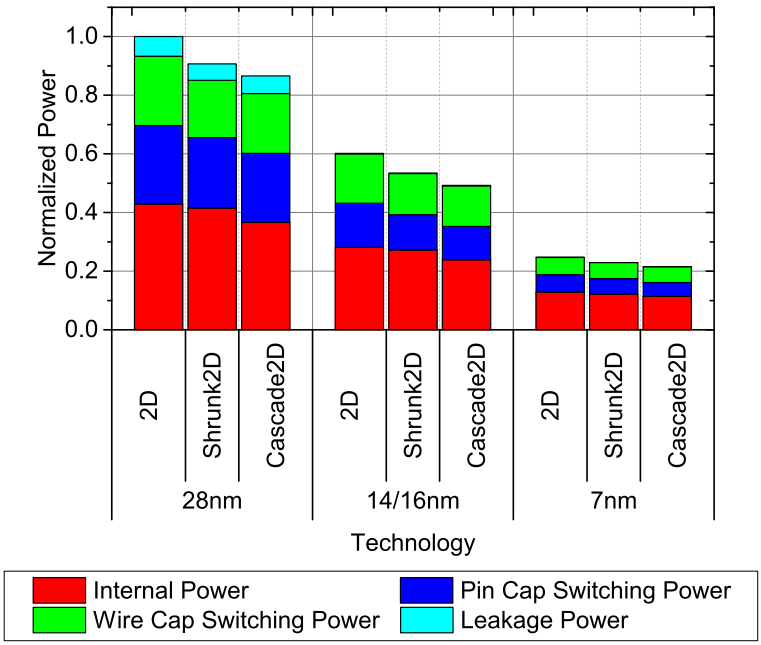

Figure 3.19: Breakdown of the power consumption of 2D, **shrunk-2D**, and **cascade-2D** M3D ICs in 28nm, 14/16nm, and 7nm technology nodes at 1.0GHz in foundry 28nm, 14/16nm, and predictive 7nm technology nodes

compared to FinFET because of the lower pin capacitance versus wire capacitance ratio at low frequencies

- With process scaling at the sub-10nm nodes, wire RC parasitic gets worse, and larger gates are required to drive increasing wire RC parasitics. Hence, pin capacitance to wire capacitance ratio is likely to increase with process scaling.
- Although power benefit from wire capacitance switching power is steady across clock frequency, it is lower compared to that achieved from standard cell area reduction
- M3D stacking technology provides maximum benefit at higher frequencies where it has the potential to reduce buffer count and drive-strengths of the cells, resulting in internal power and pin capacitance switching power reduction

3.4.2 New Monolithic 3D IC Design flow

The features of the new M3D IC design flows are summarized as follows:

- A novel M3D IC design flow, **cascade-2D design flow**, is presented which incorporates design and micro-architecture insight to guide the partitioning scheme
- **Cascade-2D design flow** is partition-scheme agnostic, hence, making it an ideal platform to evaluate different partitioning schemes
- **Cascade-2D design flow** effectively reduces standard cell area as well as wire-length compared to 2D ICs, resulting in significant power saving
- **Cascade-2D design flow** shows better power saving compared to **shrunk-2D design flow** especially in advanced technology nodes

CHAPTER 4

POWER SUPPLY INTEGRITY OF MONOLITHIC 3D ICS

4.1 Motivation and Background

Challenges in designing a reliable PDN increase mainly due to lower supply voltage, faster operating clock frequency, and higher power density. Along with restricted budget of resources and cost, these challenges may cause functional failures and performance degradation due to parasitics-induced voltage drop in a non-ideal PDN. The total voltage drop is decomposed into a resistive component (IR-drop) and an inductive component (Ldi/dt-drop). Increasing the metallization in a PDN can mitigate the resistive component of the voltage drop using wider interconnects while taking into account routing resources and cost budget.

Meanwhile, the inductance of a package including controlled collapsed chip connection (C4) bumps leads to significant Ldi/dt-drop due to time-varying current drawn by cells in a die. In order to mitigate this drop, decoupling capacitors (decaps) are utilized for local charge storage. Decaps can be placed on a die with decoupling cells (decap cells), or explicitly added in the package. However, this decap along with resistance and inductance of a PDN forms an RLC circuit resulting in its own resonance frequency [24]. If the resonance frequency lies on the system's operating frequency range, a significant Ldi/dt-drop can be induced, and hence, it is crucial to have low input impedance across a wide range of frequencies.

While the PDNs of 2D ICs (i.e., 2D PDNs) have been explored actively [24, 25], the PDNs of M3D ICs (i.e., M3D PDNs) have not been studied widely. A study for a system-level PDN for TSV-based 3D ICs is presented in [26], but the PDNs in M3D ICs and in TSV-based 3D ICs show quite different characteristics due to their tier-connection method

Table 4.1: Static and dynamic worst instance voltage drop in discrete cosine transform (DCT) 2D and M3D ICs

	2D	M3D	$\Delta\%$
static voltage drop (mV)	27.6	68.1	146.7%
dynamic voltage drop (mV)	323.4	346.9	7.3%
$\Delta\%$	1,073.9%	507.1%	-

and achievable vertical integration density. In TSV-based 3D ICs, power is delivered directly to power pads of each tiers through dedicated power TSVs, forming a parallel resistive path between multiple tiers. However, in M3D ICs, instead of having external power pads on the bottom tier, power MIVs are utilized to connect the bottommost metal layer of the top-tier PDN and the topmost metal layer of the bottom-tier PDN, consisting of a series resistive path across multiple tiers, which makes bottom-tier cells experience much longer resistive path compared to TSV-based 3D ICs. Furthermore, irregular power MIV placement due to cell blocking on the top tier makes power delivery issue more complicated in M3D ICs. For these reasons, M3D ICs suffer much higher voltage drop in the static mode, especially on the bottom-tier cells as shown in Table 4.1 compared to TSV-based 3D ICs [26].

Although the series resistive path of an M3D PDN worsens the voltage drop in the static mode, it benefits the voltage drop in the dynamic mode by improving resiliency against AC current noise, which will be discussed in later. Thus, the difference in the voltage drop between the 2D and the M3D IC in the dynamic mode is 7.3%, which is similar to TSV-based 3D ICs [26].

In [27], the authors have investigated the impact of PDNs on the power and performance of M3D ICs by proposing PDN designs, but the authors did not perform voltage drop analysis on them. An in-depth study of M3D PDNs is required to explain the significantly different trends in the voltage drop in two analysis methods, and the benefits and challenges of M3D PDNs over 2D ICs need to be investigated.

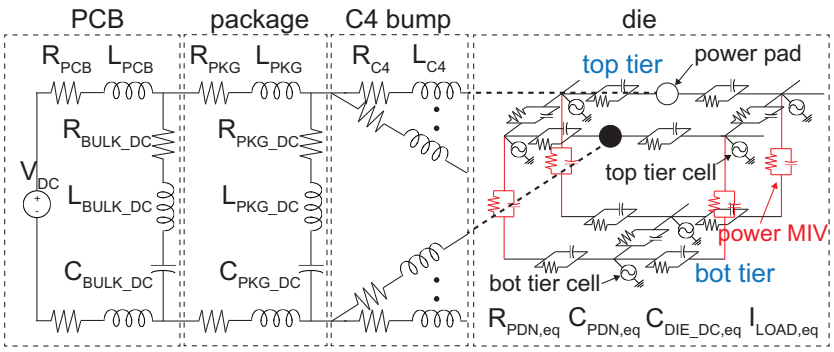

Figure 4.1: Simplified model of a system-level M3D PDN structure

4.2 System-Level Power Delivery Network Analysis and Optimization for Monolithic 3D ICs

In this work, the PDNs of 2D and M3D ICs are compared taking into account two analysis modes. The static mode is a vector-less analysis mode wherein the switching activity of cells is averaged into a single instance. In the dynamic mode, a real workload-based (i.e., vector-based) power analysis is performed for a given period of time. The dynamic mode thus incorporates the impact of inductive transients by taking into account workload-dependent time-varying current flow.

4.2.1 System-Level Power Delivery Network Modeling

In order to perform an in-depth PDN analysis, it is crucial to build a system-level PDN model. Figure 4.1 shows a simplified representation of an M3D PDN. It consists of a system model and a die model, and the system model is categorized into C4 bump, package, and printed circuit board (PCB) models [25].

In the die model, the parasitics of the PDN of a design is extracted from Cadence® Voltus™. In Figure 4.1, the resistance of the metal wire, $R_{PDN,eq}$, represents the equivalent resistive parasitics of the metal wires consisting the PDN, and the implicit decap of the die, $C_{PDN,eq}$, consists of the equivalent capacitance of the PDN metal wires, non-switching device capacitance, and coupling capacitance between N-well and substrate. The current

drawn by switched cells is lumped and modeled as an AC load current source, $I_{LOAD,eq}$.

A representative lumped system model is created based on the parameters obtained from [28] and [25]. The C4 bumps and power-line traces in the package and PCB are modeled as a series connection of the a resistor and an inductor (C4 bumps: R_{C4} = $1m\Omega$ and L_{C4} = $10pH$; package: R_{PKG} = $10m\Omega$ and L_{PKG} = $100pH$; PCB: R_{PCB} = $5m\Omega$ and L_{PCB} = $1uH$), and a DC voltage source supplies power on the PCB. The inductor and the capacitor used in the voltage regulator module (VRM) LC-tank filter are incorporated within the PCB parasitics.

Since the implicit decap alone is not sufficient to keep the design in safe voltage drop region from Ldi/dt-drop, explicit decaps are deployed both on the die using decap cells, $C_{DIE\_DC,eq}$, and on the package and PCB using discrete decaps, C_{PKG_DC} (= $400nF$) and C_{BULK_DC} (= $400uF$), respectively. The discrete decaps are modeled by a capacitor connected to an effective resistor and inductor in series (R_{PKG_DC} = $20m\Omega$ and L_{PKG_DC} = $200pH$; R_{BULK_DC} = $10m\Omega$ and L_{BULK_DC} = $2nH$). These explicit decaps and the implicit decap of the die act as charge storage elements and prevent system failure or performance degradation due to severe Ldi/dt-drop.

4.2.2 Analysis on Power Supply Integrity of Monolithic 3D ICs

Monolithic 3D IC Power Delivery Network Design Flow

Shrunk-2D design flow presented in Section 2.2.1 does not include a PDN design step, so the flow extended to incorporate a M3D PDN.

To build a full M3D PDN, before determining the location of signal MIVs in **shrunk-2D design flow**, an M3D PDN with the full metal stack (the original and the duplicated metal layers) is implemented because the signal MIVs should not be placed at the location that power MIVs are to be placed, thereby preventing the signal MIVs from deteriorating the quality of the M3D PDN. After having the PDN structure of the M3D IC including the power MIVs, the signal MIV location is determined, and the rest of steps in **shrunk-2D**

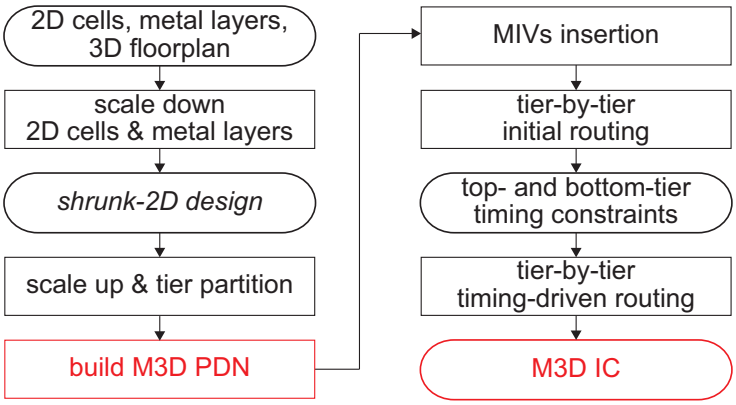

Figure 4.2: Extended **shrunk-2D design flow** to insert a PDN

design flow are performed as described in Section 2.2.1.

Once the top- and bottom-tier designs with PDNs are obtained for the final M3D IC, the designs are merged into a single M3D ICs, and timing and power analysis as well as PDN analysis is performed.

Technology Nodes and Design Libraries

Three benchmarks, DCT, AES-128, and JPEG from OpenCores are used as benchmarks in this work. NanGate FreePDK45 Open Cell Library is used to synthesize, place, and route the designs. 2D ICs and each top and bottom tier of M3D ICs are implemented using 7 metal layers. The footprint of the 2D ICs are determined such that the cell utilization is 60%, and the M3D ICs have half the footprint of the 2D counterparts. Target frequencies of each benchmarks are fixed to their maximum operating frequency in the technology node.

Table 4.2 summarizes the resources used on each metal layer to build the 2D and M3D PDNs. The dimensions of power rails are determined targeting the maximum instance IR-drop to be 5% of nominal voltage ($1.1V$) for the 2D ICs of all benchmarks in static rail analysis. For fair comparison, the same metrics are used for both the top and bottom tier of the M3D ICs. Power rails on M1 and M2 layers are tightly coupled and run in parallel in the horizontal direction of the designs, and M2 and M5 power rails are connected with only via arrays, which cross M3 and M4. M5 to M7 power rails form a mesh-structure to

Table 4.2: Width, pitch, and utilization of the 2D and M3D PDNs. Same specs are used for both 2D and M3D (both top and bottom tier) ICs.

metal layer	direction	width (μm)	pitch (μm)	utilization
M2	H	0.07	1.4	10.0%
M5	V	0.28	14	20.6%
M6	H	0.28	14	20.6%
M7	V	0.8	42	11.1%

Table 4.3: Design metrics and decoupling capacitance of the created decap cells

cell name	cell width (μm)	cell height (μm)	capacitance (fF)
DECAP_X1	0.19	1.4	3.4
DECAP_X2	0.38	1.4	6.8
DECAP_X4	0.76	1.4	13.7
DECAP_X8	1.52	1.4	27.3
DECAP_X16	3.04	1.4	54.7
DECAP_X32	6.08	1.4	109.4

distribute power across the chip.

Since NanGate FreePDK45 Open Cell Library does not provide decap cells, decap cells are created with various sizes for the experiment. Table 4.3 shows the size and decoupling capacitance of the decap cells. The decoupling capacitance of each cell is derived using the method presented in [29]. With the fully placed and routed 2D and M3D ICs, decap cells are first placed next to clock buffers driving the clock pins of flip-flops, which usually suffer from high Ldi/dt-drop. Then, rest of decap cells are placed in empty area of the designs to meet a target decoupling capacitance of the chip.

The power and ground pads of the designs are located on the top metal layer of designs (M7 for the 2D ICs, M7 of the top tier for the M3D ICs) with 120μm spacing, which model the C4 bumps of the designs.

Analysis Methods

Figure 4.3 shows the number of switched cells in the DCT design during a workload-based simulation. The vector-based power consumption in Table 4.4 is measured during the time step which shows the highest switching activity throughout the simulation (blue

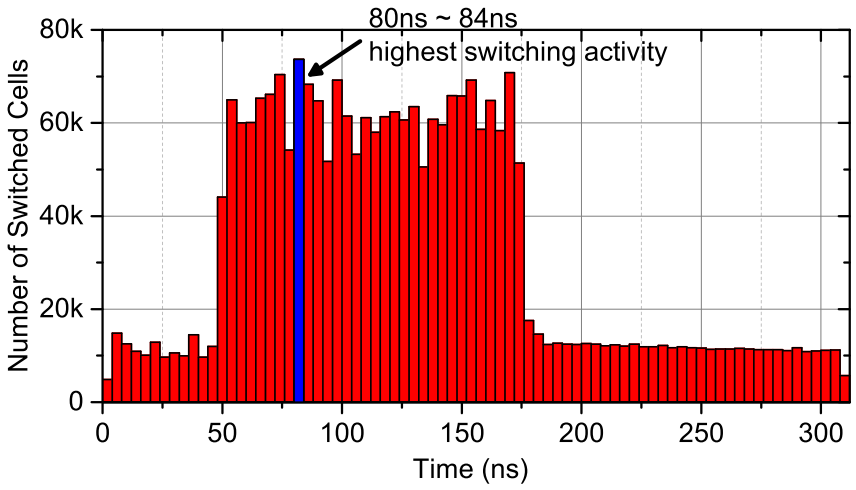

Figure 4.3: Number of the switched cells in a DCT design during a workload-based simulation. Only the time period which shows the highest switching activity (blue bar) is used for the analysis.

bar in Figure 4.3), while the statistical power consumption of the designs is calculated assuming the switching ratio of the primary input and sequential logic as 20% and 10%, respectively. Therefore, the dynamic power (i.e., internal + net switching power) shows significant difference between two analysis methods whereas the static power (i.e., leakage power) remains similar.

The M3D ICs offer power benefit over their 2D counterparts. Since M3D ICs utilize short vertical integration with MIVs instead of using long metal wires on the xy-plane, the wire-length of the designs are reduced as shown in Table 4.4, offering net switching power saving. In addition, since the cells drive the reduced wire-load, the number of buffers as well as the drive-strength of the cells decreases, which, in turn, reduces the standard cell area, hence, showing benefits on the internal and leakage power consumption.

Instance voltage drop is used, which is the voltage drop a cell experiences as described in Equation (4.1).

$$\Delta V_{inst} = (V_{DD,nom} - V_{DD,act}) + (V_{SS,act} - 0), \tag{4.1}$$

where ΔV_{inst} is instance voltage drop, and $V_{DD,nom}$ is nominal voltage. $V_{DD,act}$ and $V_{SS,act}$ are actual voltage level on the power and ground pin of the cell. Instance voltage drop

Table 4.4: Iso-performance design and power metric comparison of 2D and M3D ICs used for static and dynamic rail analysis. Both statistical and vector-based power simulations are conducted. $\Delta\%$ for M3D ICs is calculated with respect to the 2D counterparts.

benchmark		DCT			AES-128			JPEG		
		2D	M3D	$\Delta\%$	2D	M3D	$\Delta\%$	2D	M3D	$\Delta\%$
frequency (MHz)		500	500	-	1,000	1,000	-	500	500	-
footprint (μm)		369×368	260×260	-50.2%	509×507	360×360	-49.8%	897×895	634×636	-49.8%
C4 bump count		9	4	-55.6%	16	9	-43.8%	64	25	-60.9%
std cell area (μm^2)		85,432	85,312	-0.1%	166,560	163,938	-1.6%	503,070	503,068	0.0%
wire-length (μm)		784,072	723,139	-7.8%	1,921,276	1,708,362	-11.1%	3,770,356	3,730,150	-1.1%
total capacitance (pF)		236.1	220.0	-6.8%	568.6	500.3	-12%	1,186.4	1,153.7	-2.8%
signal MIV count		-	11,753	-	-	50,589	-	-	58,807	-
statistical power analysis	internal power (mW)	16.4	16.2	-1.6%	49.3	48.5	-1.6%	109.4	109.8	0.4%
	net switching power (mW)	15.9	13.4	-15.8%	46.8	41.1	-12.3%	93.4	89.3	-4.4%
	leakage power (mW)	0.8	0.7	-2.8%	1.9	1.7	-7.6%	4.5	4.4	-1.9%
	total power (mW)	33.1	30.3	-8.4%	98.0	91.3	-6.8%	207.3	203.6	-1.8%
vector-based power analysis	internal power (mW)	48.0	51.2	6.7%	197.3	186.0	-5.7%	218.1	222.1	1.8%
	net switching power (mW)	36.7	31.3	-14.7%	134.8	87.9	-34.8%	88.8	91.8	3.4%
	leakage power (mW)	0.7	0.7	-3.0%	2.0	1.8	-7.2%	4.7	4.6	-2.1%
	total power (mW)	85.4	83.2	-2.6%	334.0	275.7	-17.5%	311.6	318.5	2.2%

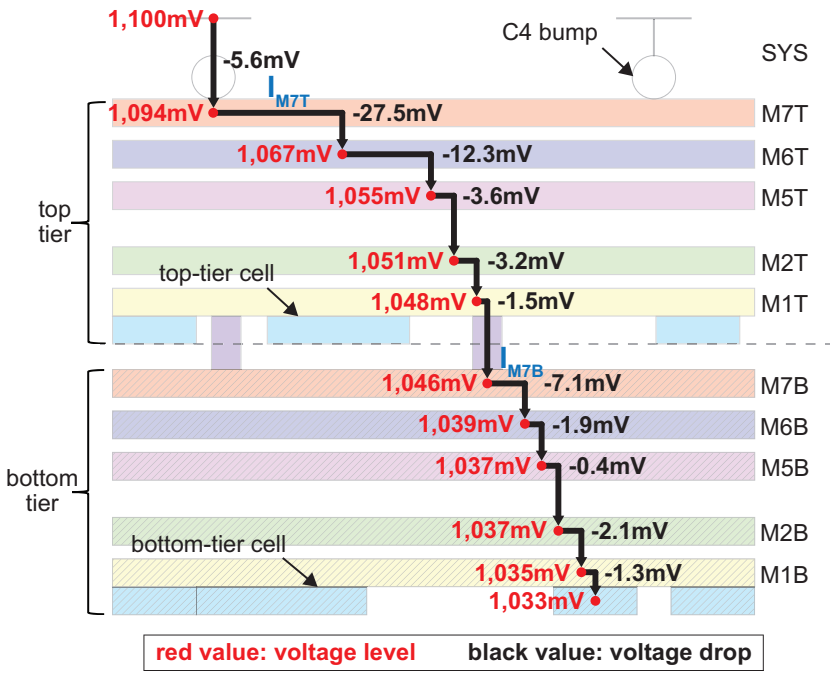

Figure 4.4: Illustration describing how the worst instance IR-drop can be decomposed into each metal layer, showing voltage drops on the power rails in each metal layer (values in black) and voltage level on each metal layer along the IR-drop path (values in red).

can be further decomposed into the voltage drops on each metal layer. Figure 4.4 shows the decomposed voltage drop at each metal layer (voltage values in black) for the cell experiencing the worst instance voltage drop in static rail analysis of the JPEG M3D IC, showing how much IR-drop the power rails in each metal layer has contributed to the total instance IR-drop.

Static Rail Analysis

Since static rail analysis is based on statistical power consumption, which summarizes the behavior of designs, only IR-drop can be analyzed. Figure 4.5 shows the breakdown of the worst instance IR-drop into each metal layers consisting the PDN of the 2D and M3D ICs. The M3D ICs show approximately 2× higher IR-drop on average compared to the 2D counterparts. Since M3D ICs utilize more metal layers for their PDN structure to deliver power to the bottom-tier cells, those cells experience worse IR-drop compared to the top-tier cells (The dashed box from M1B to M7B in Figure 4.5).

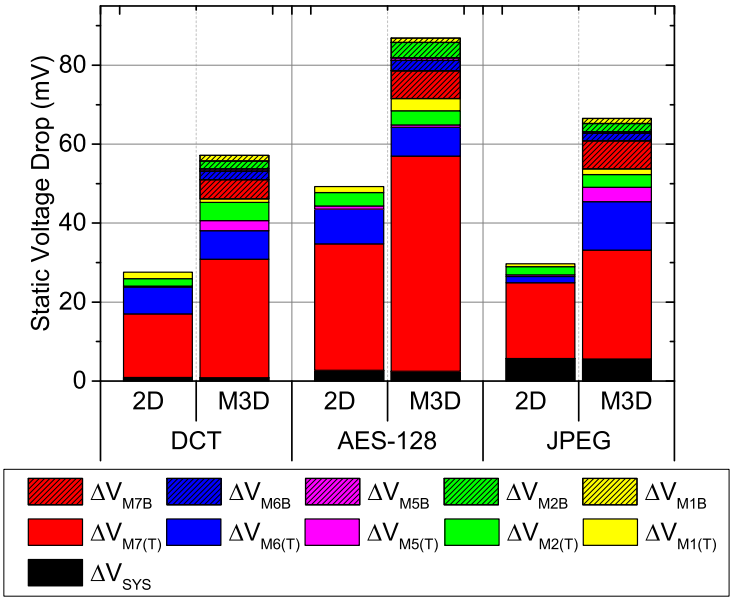

Figure 4.5: Breakdown of the worst instance IR-drop across the metal layers comparing 2D and M3D ICs. M7B denotes M7 of the bottom tier in the M3D ICs. SYS represents the system model including C4 bump, package, and PCB model.

Another reason for the higher IR-drop in M3D ICs is irregular placement of power MIVs which connect PDNs of two tiers. Figure 4.6 illustrates the impact of irregular power MIV placement. In M3D PDNs, current flowing in metal layers of the top tier is greater than those of the bottom tier (e.g., $I_{M7T} > I_{M7B}$ in Figure 4.6) since top-tier metal layers deliver current to both top- and bottom-tier cells whereas only current drawn by bottom-tier cells flows on bottom-tier metal layers. Therefore, the minimum IR-drop path in Figure 4.6 to deliver power to a bottom-tier cell utilizes the minimum length of top-tier power rails. However, the path can be blocked by a missing power MIV. The absence of power MIVs stems from top-tier cells since MIVs cannot penetrate those cells in order to preserve their active areas. In this case, the current needs to flow through an alternative path shown as actual path in Figure 4.6, which utilizes longer top-tier metal wires and hence, exhibits worse IR-drop due to higher current in those wires. It is important to note that top-tier metal layers are more susceptible to electromigration because of higher currents, however that discussion is out of the scope of this work.

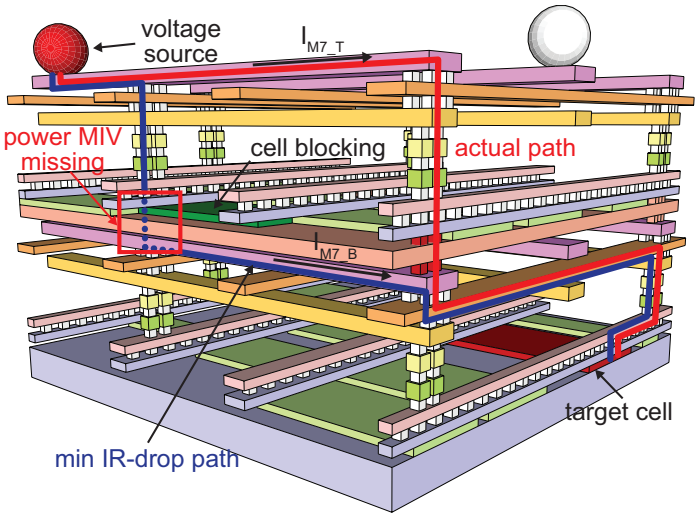

Figure 4.6: Current path to deliver power to a target cell showing the impact of missing power MIVs. A top-tier cell is blocking a power MIV along the minimum IR-drop path, so the current is delivered through an alternative path.

Table 4.5: Average amount of current flowing through C4 bumps in 2D and M3D ICs

design	2D	M3D	Δ%
DCT	3.34mA	6.89mA	106.1%
AES-128	5.48mA	9.23mA	68.2%
JPEG	2.87mA	7.34mA	155.7%

Reduced number of C4 bumps in M3D ICs also degrades voltage integrity. As the footprint of an M3D IC is half of its 2D counterpart, the number of the C4 bumps that can be placed on an M3D IC is approximately half of those in the 2D IC as shown in Table 4.4. This affects the amount of the current flowing through each C4 bump. Table 4.5 compares the current flowing through C4 bumps in the 2D and M3D ICs. Up to 155.7% higher current flows through the C4 bumps in the M3D ICs incurring significant difference in IR-drop on the top metal layers (i.e., M7 and M6 in Figure 4.5).

Dynamic Rail Analysis

Unlike static rail analysis, dynamic voltage drop consists of two categories, IR-drop and L$d i$/dt-drop. L$d i$/dt-drop has significantly higher impact on the voltage drop since dynamic rail analysis is performed for two clock cycles with the maximum switching activity in a

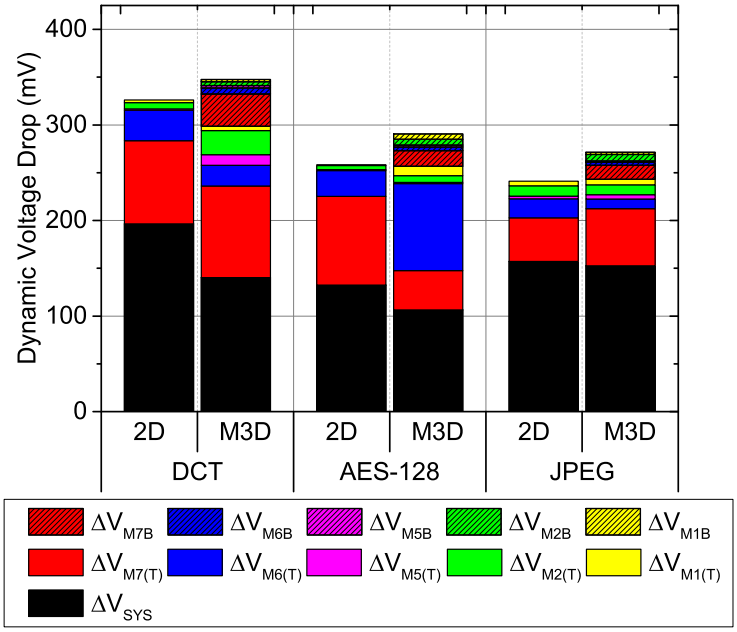

Figure 4.7: Breakdown of the worst instance dynamic voltage drop (= IR-drop + Ldi/dt-drop) across the metal layers comparing 2D and M3D ICs

real workload. The voltage drop of the M3D ICs is 11.3% higher on average than the 2D ICs as shown in Figure 4.7, which is much less than that in the static rail analysis.

First, in the dynamic rail analysis, the difference between the voltage drop of the metal layers in the 2D ICs and the metal layers of the top-tier PDN of M3D ICs is significantly less than that in the static rail analysis. The reduced voltage drop on those metal layers first results from 3D placement of decaps in M3D ICs. As discussed in Section 4.2.1, decaps from non-switching devices (implicit) and decap cells (explicit) in a design act as charge reservoir, preventing nearby cells from experiencing sudden high Ldi/dt-drop. In M3D ICs, Ldi/dt-drop is reduced because of decap cells in both, the xy-plane, as in the case of 2D ICs, as well as decap cells in the adjacent tier (the z-axis), as in TSV-based 3D ICs [26].

Figure 4.8 shows the maximum voltage drop experienced at the C4 bumps comparing the DCT 2D and M3D ICs with and without decap cells. The decoupling capacitance of the 2D and M3D ICs with decap cells is targeted to 30% of their total capacitance. Even though the decoupling capacitance added to the M3D IC (= 72.6pF) is smaller than the 2D

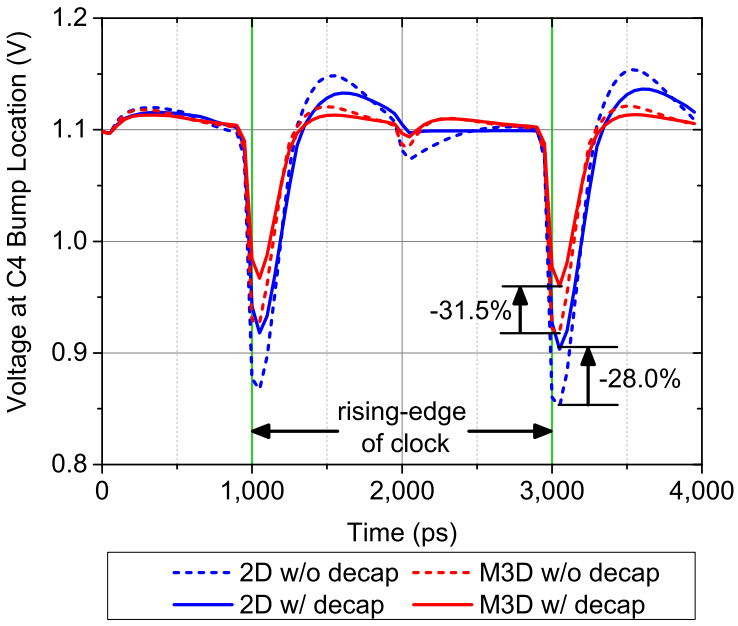

Figure 4.8: Comparison of the worst voltage drop experienced at C4 bumps showing the impact of decoupling capacitance in 2D and M3D ICs. The decoupling capacitance is set to 30% of the total capacitance of each design.

IC (= 77.9pF) due to the lower total capacitance of the DCT M3D IC, it benefits more from the added decap cells than the 2D IC as it utilizes decaps in the z-axis as well. Another reason for the smaller gap between the 2D and M3D IC voltage drop in the dynamic rail analysis is the reduced voltage drop on the system model as shown in Figure 4.7 due to the varying impedance seen from the die depending on operating frequency, which will be discussed in the next sub-section.

Frequency- and Time-Domain Analysis

As shown in Figure 4.1, implicit and explicit decaps on a die model are coupled with inductors in a system model, forming an RLC circuit. The RLC circuit has its own resonance frequency, causing significant voltage drop on the PDN even with small changes in load current. Explicit decaps on the package and PCB also forms RLC circuits with the corresponding inductors, showing their unique resonance frequencies.

To perform an in-depth frequency- and time-domain analysis on a PDN, a reasonable

Table 4.6: Effective resistance and capacitance of 2D and M3D PDNs. Resistance is in Ω, and capacitance in nF.

parameter		2D	M3D	Δ%
DCT	$R_{PDN,eq}$	0.221	0.812	267.5%
	$C_{PDN,eq}$ + $C_{DIE\_DC,eq}$	0.232	0.211	-8.7%
	product of R and C	0.051	0.172	235.5%
AES-128	$R_{PDN,eq}$	0.164	0.341	108.1%
	$C_{PDN,eq}$ + $C_{DIE\_DC,eq}$	0.528	0.439	-16.8%
	product of R and C	0.086	0.150	73.1%
JPEG	$R_{PDN,eq}$	0.076	0.202	164.4%
	$C_{PDN,eq}$ + $C_{DIE\_DC,eq}$	1.290	1.220	-5.4%
	product of R and C	0.098	0.246	150.1%
average product of R and C		0.079	0.189	140.4%

die model which represents 2D and M3D full-chip System-on-Chip (SoC) is needed. Since the benchmarks used in this work are small compared to full-chip SoCs, their parameters are used to create a full-chip die model. Table 4.6 shows the effective resistance and capacitance of the PDN of each benchmark ($R_{PDN,eq}$ and $C_{PDN,eq}$ + $C_{DIE\_DC,eq}$ in Figure 4.1, respectively). As a design becomes larger, the capacitance of its PDN increases due to the increased ground and coupling capacitance of the PDN, while the resistance becomes smaller because more number of parallel resistive paths to the cells are available. For ease in modeling, the average of the RC product from the three benchmarks is used, and a full-chip die is modeled by assuming $C_{PDN,eq}$ + $C_{DIE\_DC,eq}$ = $10nF$, resulting in the associated resistances as $7.87m\Omega$ and $18.9m\Omega$ for the 2D and M3D ICs, respectively.

Figure 4.9 shows the frequency response of the 2D and M3D full-chip SoC sweeping the frequency of the AC load current source, $I_{LOAD,eq}$. Three resonance frequency points are observed, first-order resonance caused by $C_{PDN,eq}$ + $C_{DIE\_DC,eq}$ coupled with L_{C4}, second-order resonance by C_{PKG_DC} with L_{PKG}, and third-order resonance by C_{BULK_DC} with L_{PCB}. While third-order and second-order resonance occurs at a few kHz and MHz range, the largest resonance, first-order resonance is in the range between $50MHz\sim200MHz$. Although the M3D IC shows 16.7% increase at second-order resonance frequency, as the operating frequencies of full-chip SoC at advanced technology nodes are in the range of

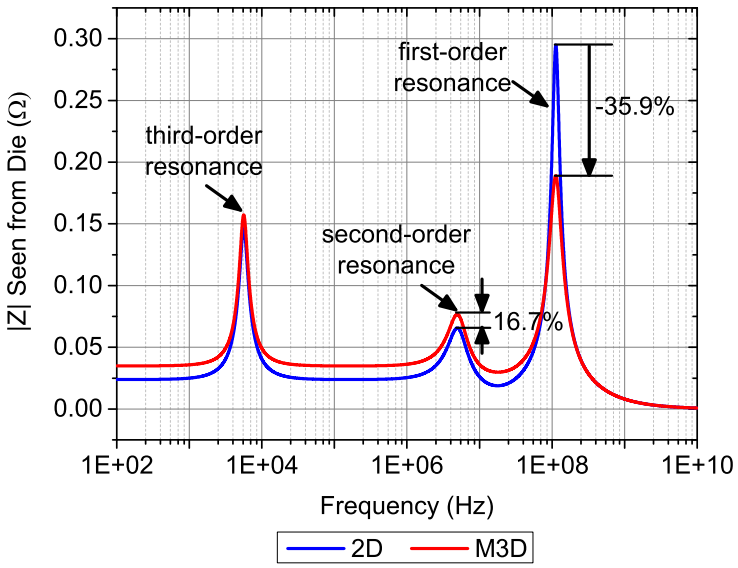

Figure 4.9: Impedance seen from the die by sweeping the frequency of AC load current source, $I_{LOAD,eq}$

first-order resonance frequencies, it is crucial to minimize the first-order resonance impact for a robust PDN.

As shown in the figure, the M3D IC exhibits 35.9% lower peak impedance at first-order resonance frequency because of high effective resistance of the M3D PDN due to the series resistive path across tiers. An interesting point is that the high resistance of M3D PDNs, which worsens IR-drop, in fact, improves the resiliency against AC current noise by damping noise at worst-case resonance oscillation. This work is the first study to demonstrate this effect of M3D ICs.

Figure 4.10 (a) and (b) shows the improved resiliency of the M3D PDN, showing the time-domain response for a unit step, which models in-rush current simulation, and for a $112MHz$ (first-order resonance frequency) unit sine-wave load current source. Equation (4.2) explains the die voltage response affected by first-order resonance for a unit step load current source [25].

$$\Delta V_{DIE} \cong 2R + \sqrt{\frac{2L_{C4}}{C_{DIE,eq}}} \cdot e^{-\frac{R}{2L_{C4}}t} sin(\omega_r - \theta), \tag{4.2}$$

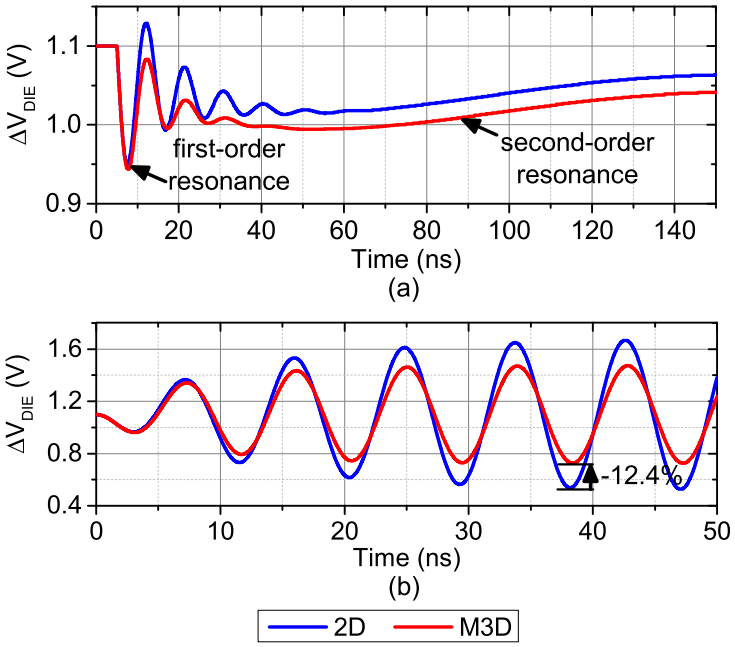

Figure 4.10: Transient voltage response for (a) a unit step, and (b) a unit $112MHz$ (first-order resonance frequency) sine-wave load current source, $I_{LOAD,eq}$. Third-order resonance frequency is not shown in (a) for brevity.

where R = R_{PCB} + R_{PKG} + R_{C4} + $R_{PDN,eq}$, ω_r and θ are first-order resonance frequency and phase, respectively. While the increased R in an M3D PDN worsens the IR-drop at cell (the first term in Equation (4.2)), it helps to reduce the second term, Ldi/dt-drop. The improved resiliency for first-order resonance helps neutralize the voltage drop gap induced by second-order resonance in the worst voltage drop as shown in Figure 4.10 (a), and shows 12.4% less voltage drop with current source oscillating at first-order resonance frequency as shown in Figure 4.10 (b).

4.2.3 Power Delivery Network Optimization Methodologies for Monolithic 3D ICs

Although 3D placement of decap cells and better resiliency against AC current noise help reduce Ldi/dt-drop, M3D ICs still suffer from higher IR-drop due to additional metal layers, irregular placement of power MIVs, and fewer C4 bumps. In addition, as technology advances, IR-drop of M3D ICs becomes worsened because of increased PDN resistance due to restricted routing resources, whereas the increased resistance offer better resiliency

against AC current noise, especially in M3D ICs, mitigating Ldi/dt-drop. Therefore, the optimization methodologies mainly focus on reducing the IR-drop of M3D ICs by repositioning top-tier cells and utilizing asymmetric top- and bottom-tier M3D PDNs.

Top-Tier Cell Repositioning

Among the factors of the worse IR-drop of M3D ICs, additional metal layers and fewer C4 bumps are inevitable due to the nature of M3D ICs. However, irregular power MIV placement is avoidable by preventing top-tier cells from being placed at the location that power MIVs are supposed to lie in.

To achieve this, after performing the area-balanced min-cut tier partitioning in the baseline M3D IC design flow presented in Section 4.2.2, the top-tier cells, which are located above the topmost power rails of the bottom tier, are repositioned to the nearest available space. This produces empty spaces for power MIVs, making them tightly couple top- and bottom-tier PDNs without any obstruction due to the top-tier cells. Figure 4.11 (b) shows the repositioned top-tier cells, which reserve rooms for power MIVs to connect the power rails of the bottommost power rails of the top tier to the topmost power rails of the bottom tier without any obstruction as opposed to Figure 4.11 (a).

As discussed in Section 4.2.2, to minimize IR-drop of M3D ICs, the IR-drop path from current sources (i.e., C4 bumps) to a bottom-tier cell should utilize the minimum top-tier power rails because of the large current flowing through the top-tier power rails. Top-tier cell repositioning technique ensures power MIVs to be placed every intersection of the topmost power rails of the bottom tier and the bottommost power rails of the top tier, so that helps current flow through the optimal IR-drop path. Figure 4.12 shows the impact of top-tier cell repositioning by comparing the same IR-drop path (i.e., the IR-drop path to the worst instance IR-drop cell in the baseline M3D IC (i.e., M3D-base)). While M3D-base in Figure 4.12 (a) utilizes long top-tier power rails (red/blue dotted lines), the M3D IC with top-tier cell repositioning technique (i.e., M3D w/ RP) in Figure 4.12 (b) utilizes shorter

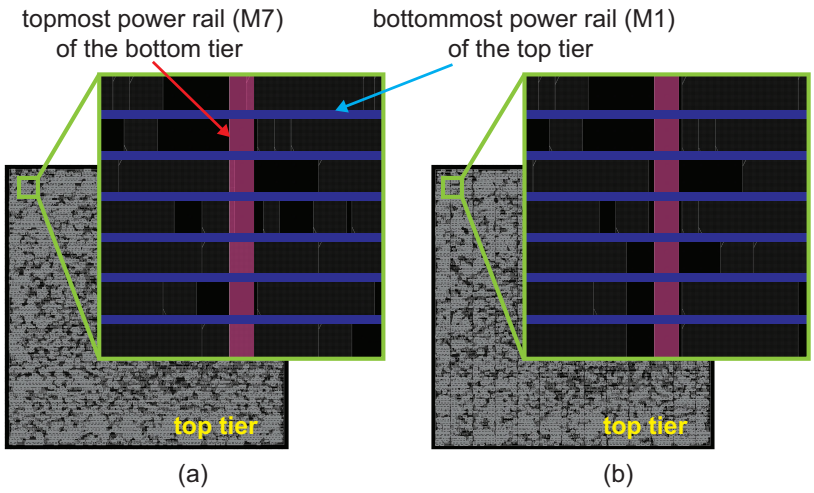

Figure 4.11: Cell placement of top-tier cells (a) in the baseline M3D (M3D-base) and (b) in M3D with top-tier cell repositioning technique (M3D w/ RP) which repositions obstructing top-tier cells to the nearest available space.

top-tier power rails, mainly using bottom-tier power rails (red/blue solid lines) and hence, mitigating the IR-drop of the design.

Table 4.7 shows the benefit of top-tier cell repositioning technique on the static voltage drop of DCT, AES-128, and JPEG M3D ICs. With repositioning top-tier cells, static voltage drop is improved up to 14.8%. To analyze this trend more in detail, the breakdown of the static voltage drop across metal layers is presented in Figure 4.13. Although top-tier cell repositioning technique tends to increase the IR-drop of the bottom-tier power rails (the dashed box from M1B to M7B; at most 14.8% in the AES-128 design) as it makes the IR-drop path utilize more bottom-tier power rails, it efficiently reduces the IR-drop on the top-tier power rails (up to 21.6% in the AES-128 design). Considering the IR-drop of the top-tier power rails dominates the total IR-drop, the technique significantly reduce the total static voltage drop of M3D ICs.

However, top-tier cell repositioning technique might degrade the timing integrity and dynamic voltage drop, which result from the increased parasitics and power consumption as shown in Table 4.7. These increases are mainly attributed to the wire-length increase, which results from top-tier cell movement. In the M3D IC design flow used in this work,

Table 4.7: Key metric comparison of the baseline M3D (M3D-base) and M3D ICs with top-tier cell repositioning technique (M3D w/ RP). $\Delta\%$ for M3D w/ RP designs is calculated with respect to the respective M3D-base designs.

benchmark	DCT			AES-128			JPEG		
	M3D-base	M3D w/ RP	$\Delta\%$	M3D-base	M3D w/ RP	$\Delta\%$	M3D-base	M3D w/ RP	$\Delta\%$
clock period (ps)	2,000			1,000			2,000		
wire-length (μm)	723,139	732,467	1.3%	1,708,362	1,729,693	1.2%	3,730,150	3,765,106	0.9%
WNS (ps)	17	4.2	-75.3%	23.9	19.1	-20.1%	18.1	9.7	-46.4%
static power (mW)	30.3	30.3	0.0%	91.3	91.4	0.1%	203.6	204.0	0.2%
static voltage drop (mV)	57.6	49.1	-14.8%	86.1	74.4	-13.6%	68.2	63.5	-6.9%
dynamic power (mW)	83.2	87.7	5.5%	275.7	277.5	0.7%	318.5	328.1	3.0%
dynamic voltage drop (mV)	352.8	369.8	4.8%	290.6	271.4	-6.6%	272.3	273.4	0.4%

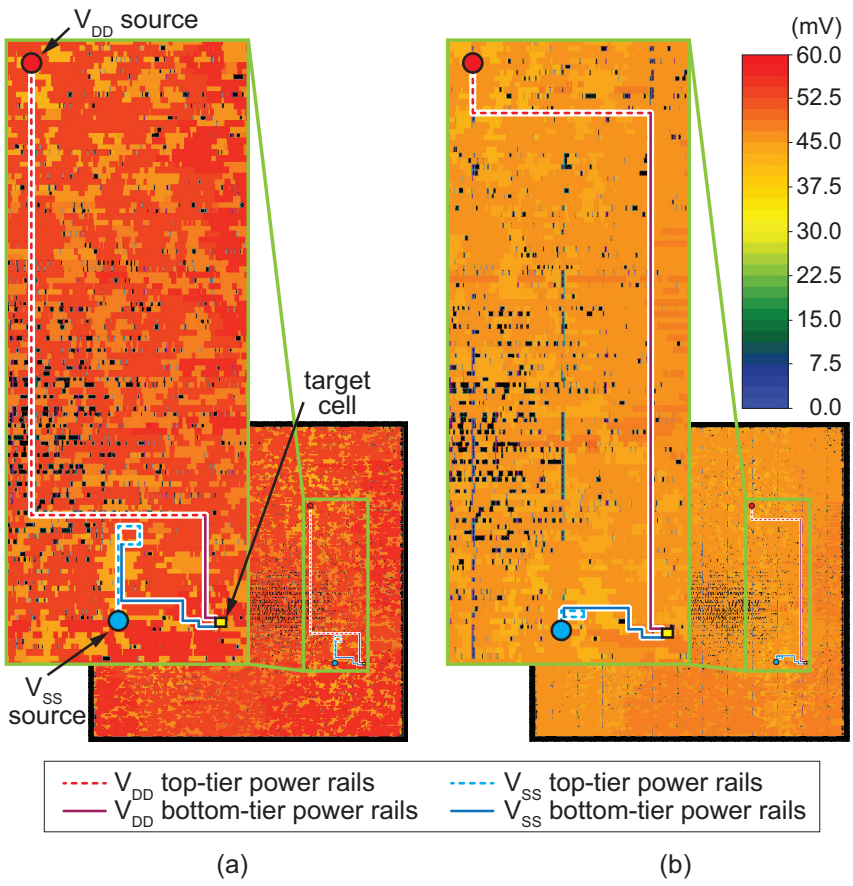

Figure 4.12: Comparison of the IR-drop path of (a) M3D-base and (b) M3D w/ RP. Red/blue lines represent the IR-drop paths of power/ground, and dotted/solid lines represent the top- and bottom-tier power rails, respectively.

cell placement is determined and optimized during implementing a *shrunk2D design*, so that the wire-length among cells are minimized. However, while top-tier cells are moved to preserve rooms for power MIVs, the wire-length is increased. As the increase in wire-length is rather minor (up to 1.3% in the DCT design), although the technique causes degradation in slack of the critical timing path (WNS), but not significant compared with the clock period. However, it significantly increases the peak dynamic power consumption (up to 5.5% in the DCT design), especially when the lengthened wires are heavily used in the vector-based power simulation.

This degradation in dynamic voltage drop can be resolved by using asymmetric top- and bottom-tier M3D PDNs, which will be discussed in the next sub-section.

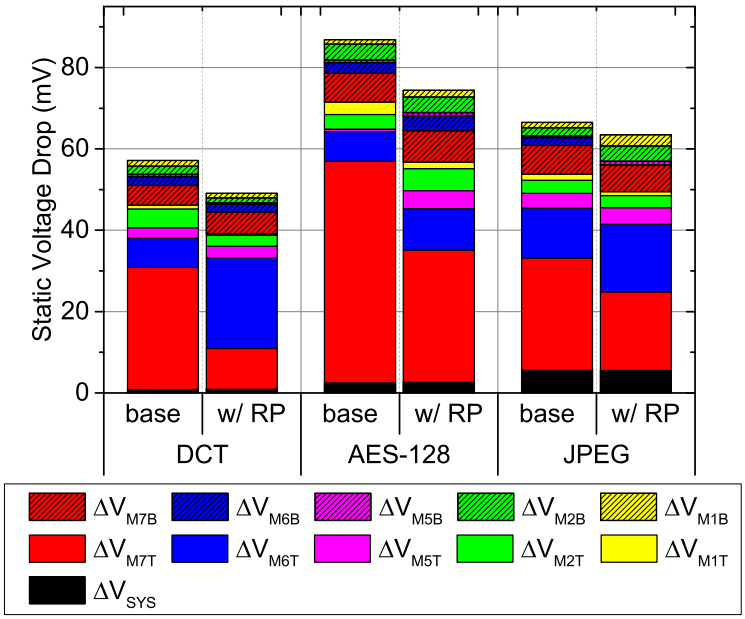

Figure 4.13: Comparison of the breakdown of the worst instance IR-drop across metal layers between M3D-base (base) and M3D w/ RP (w/ RP)

Asymmetric Top- and Bottom-Tier PDN

Although top-tier cell repositioning technique helps reduce the IR-drop on top-tier power rails, the top-tier IR-drop still takes large portion of the total IR-drop of M3D ICs as shown in Figure 4.13. As the large current flowing through top-tier power rails is unavoidable because they deliver power to both top- and bottom-tier cells, the IR-drop should be reduced by decreasing the resistance of top-tier power rails with an asymmetric top- and bottom-tier PDN.

Figure 4.14 shows the cross-sectional view of an asymmetric top- and bottom-tier M3D PDN. The width of top-tier power rails are increased by an unbalancing factor, α, to decrease the resistance and accommodate large current which deliver power to both top- and bottom-tier cells. On the other hand, the width of bottom-tier power rails are decreased by the same rate, α, since rather small current, which is drawn by only bottom-tier cells, flows on the bottom-tier power rails. This helps retain the signal routing resource taken by the

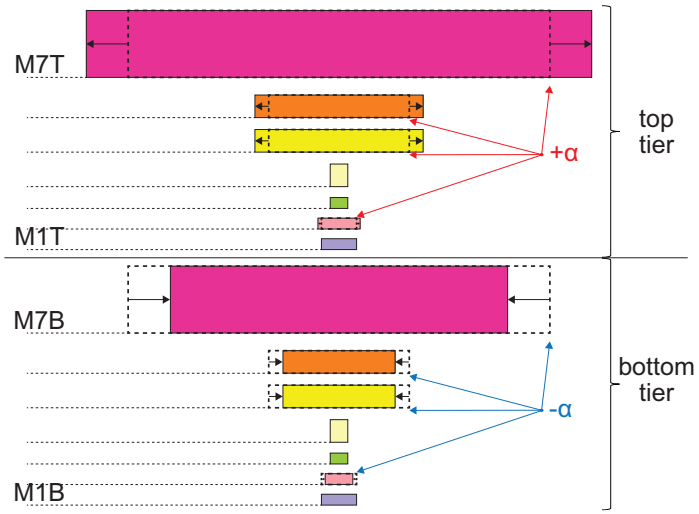

Figure 4.14: Cross-sectional view showing an asymmetric top- and bottom-tier M3D PDN. Whereas the width of top-tier power rails are scaled up by α, that of bottom-tier power rails are scaled down by α. Note that the width of M1 is not scaled as it is determined by power-pin width of standard cells, and M3 and M4 is not used for top- and bottom-tier power rails in the designs.

wider top-tier power rails because narrower power rails occupy smaller number of signal routing tracks. The width of M1 of both top- and bottom-tier power rails is not scaled since it is determined by power and ground pins of standard cells.

The worst instance IR-drop with different unbalancing factors from 10% to 50% for M3D-base and M3D w/ RP is presented in Figure 4.15. As the unbalancing factor increases, static voltage drop tends to decrease, showing only 25.7% higher static voltage drop in the AES-128 M3D IC compared to its 2D counterpart. To analyze the trend, Equation (4.3) is employed which determines the optimal unbalancing factor assuming the current flowing through the top-tier power rails is exactly twice of that through the bottom-tier power rails along the worst instance IR-drop path.

$$\begin{aligned} min(\Delta V_{DIE}) &= min(I_{top} \cdot R_{top} + I_{bot} \cdot R_{bot}) \\ &= min(2 \cdot I_{bot} \cdot (\rho \cdot l/(t \cdot w(1+\alpha))) + I_{bot} \cdot (\rho \cdot l/(t \cdot w(1-\alpha)))) \\ &= 2.91 \cdot I_{bot} \cdot (\rho \cdot l/(t \cdot w)) \text{ at } \alpha = 0.17, \end{aligned} \tag{4.3}$$

where, I_{top} and I_{bot} are the current flowing the top- and bottom-tier power rails, and ρ is the resistivity of the power rails. l, t, and w are the length, thickness and width of the power

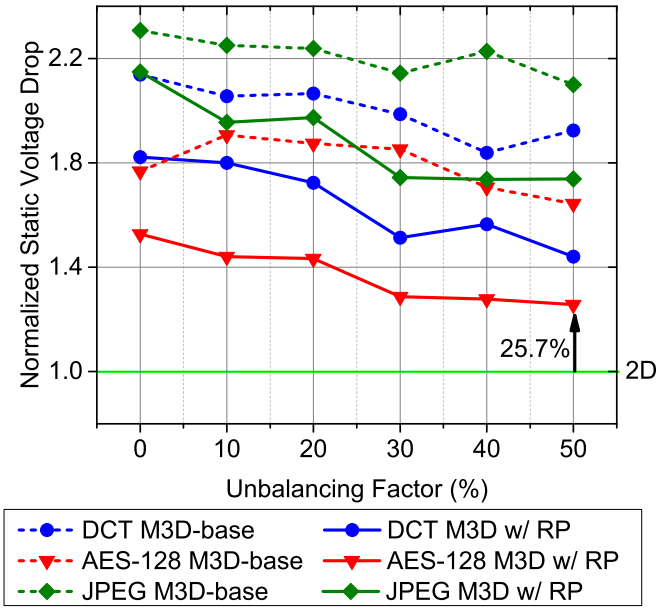

Figure 4.15: Impact of asymmetric top- and bottom-tier PDN technique on static voltage drop in M3D-base and M3D w/ RP comparing different unbalancing factors. Static voltage drop values are normalized to their 2D counterparts.

rails, respectively, and α is the unbalancing factor.

The equation indicates that the static voltage drop is minimized when the width of top-tier power rails are scaled up by 17%, and bottom-tier power rails are scaled down by 17%, which is different from the experiment result. This difference first comes from the fact that the current flowing through top-tier power rails is not exactly twice of bottom-tier power rails.

In addition and most importantly, the current sources of the top and bottom tier affect the difference. Current is delivered to bottom-tier power rails through power MIVs whose pitch is in nm scale, which establish a large number of current sources for the bottom tier. This significantly reduces current flowing through the power rails near power MIVs, hence, reducing the IR-drop of top metal layers of the bottom tier. On the other hand, the current sources of the top tier are C4 bumps whose pitches are in μm scale. Therefore, there are only a few C4 bumps in a design as shown in Table 4.4. For that reason, there is significantly large current flowing near C4 bumps, which worsens IR-drop of the design as

shown in large IR-drop on top metal layers of the top tier (i.e., M7T and M6T in Figure 4.5). In order to compensate the large current near C4 bumps on the top tier, wider width than the ratio derived from Equation (4.3) is required for top-tier power rails.

Another interesting point in Figure 4.15 is that in M3D-base, the decrease in the static voltage drop with the increasing unbalancing factor is rather smaller and not monotonic. Figure 4.16 shows the impact of irregular power MIV placement on the asymmetric top- and bottom-tier M3D PDN in the AES-128 designs. Whereas the IR-drop of the top-tier PDN is monotonically decreasing in both M3D-base and M3D w/ RP designs as the unbalancing factor increases, the IR-drop of the bottom-tier PDN tends to increase since the resistance increases as their width is scaled down. However, in M3D-base, the irregularly placed power MIVs increases the current flowing power MIVs as the number of power MIVs are smaller than M3D w/ RP, significantly increasing the IR-drop of top metal layers of bottom-tier power rails. Furthermore, as current flows through a longer alternative path which has higher resistance rather than the shorter optimal path, it also increases the bottom-tier IR-drop.

These trends also apply to dynamic voltage drop. Figure 4.17 presents the normalized dynamic voltage drop of M3D ICs changing with different unbalancing factors, even showing less dynamic voltage drop in the AES-128 M3D ICs compared to their 2D counterparts. Static voltage drop reduction due to the asymmetric top- and bottom-tier M3D PDN in conjunction with higher resiliency against AC current noise due to higher M3D PDN resistance helps reduce dynamic voltage drop of M3D ICs.

However, the widened top-tier power rails negatively affects the timing of M3D ICs, but the impact can be minimized by utilizing more bottom-tier routing resources for signal routing. Figure 4.18 shows an example of connecting two top-tier cells, when the top-tier routing resource is restricted due to the increased width of the top-tier power rails. Instead of routing them using only top-tier metal layers (dashed wires in Figure 4.18), it utilizes signal MIVs to connect the two top-tier cells through bottom-tier metal layers (solid wires

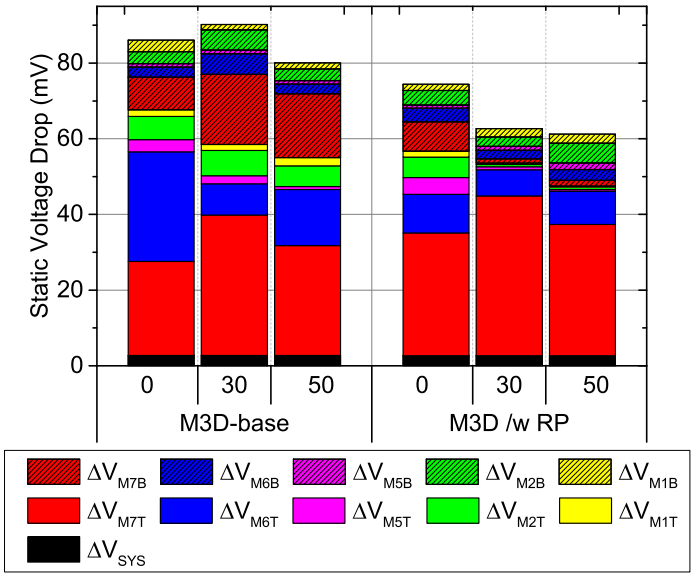

Figure 4.16: Breakdown of the worst instance IR-drop across the metal layers with 0%, 30%, and 50% unbalancing factors in asymmetric top- and bottom-tier PDN technique comparing AES-128 M3D-base and M3D w/ RP

in Figure 4.18).

Table 4.8 compares the design and timing metrics of the AES-128 M3D ICs with unbalancing factor 0% and 50%. As shown in the table, M3D IC with unbalancing factor 50% utilizes less wires on the top tier because of the reduced available resources caused by the widened top-tier power rails. Instead, it uses more routing resources of the bottom tier and more signal MIVs because of the increased available routing resources on the bottom tier.

As the M3D IC with the unbalancing factor 50% utilizes more bottom tier wires, wire ground and coupling capacitance of the bottom tier are increased, while those of the top tier are decreased, keeping the total wire capacitance almost the same (only $0.4pF$ increase), and hence, the impact on the critical path is quite low, considering the clock period of the design.

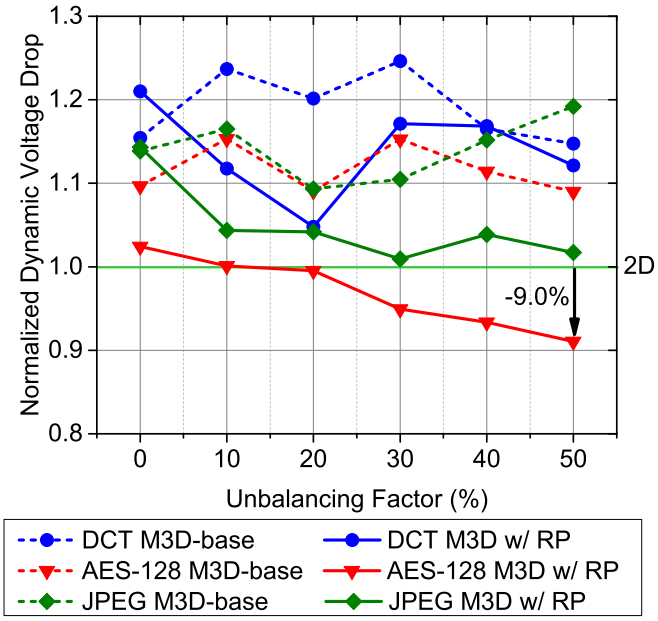

Figure 4.17: Impact of asymmetric top- and bottom-tier PDN technique on dynamic voltage drop in M3D-base and M3D w/ RP comparing different unbalancing factors

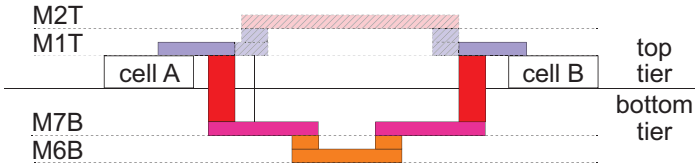

Figure 4.18: An example of using bottom-tier routing resources when top-tier metal layers are congested

4.3 Summary

4.3.1 Resistive Voltage Drop in Monolithic 3D ICs

The observations from the static rail analysis on M3D ICs are summarized as follows:

- IR-drop in M3D ICs is approximately 2× higher IR-drop than 2D ICs in the static rail analysis based on the statistical power simulation
- Increased resistive path due to additional metal layers to pass through to supply power to bottom-tier cells increases IR-drop of M3D ICs
- Irregular placement of power MIVs due to top-tier cells prevents current from flowing

Table 4.8: Key metric comparison of the M3D IC with the unbalancing factor 0% and 50%. Δ% for the unbalancing factor, α = 50% design is calculated with respect to α = 0% design.

unbalancing factor		0%	50%	
clock period (ps)		1,000	1,000	(0.0%)
bot tier	wire-length (μm)	824,715	836,295	(1.4%)
	ground cap (pF)	63.8	64.5	(1.1%)
	coupling cap (pF)	56.9	57.8	(1.6%)
top tier	wire-length (μm)	883,647	873,619	(-1.1%)
	ground cap (pF)	65.4	64.7	(-1.1%)
	coupling cap (pF)	62.2	61.7	(-0.8%)
signal MIV count		50,589	51,127	(1.1%)
WNS (ps)		23.9	14.3	(-40.2%)

the optimal resistive path

- High current flowing C4 bumps due to reduced number of C4 bumps placed in a M3D IC negatively impacts the IR-drop of M3D ICs

4.3.2 Inductive Voltage Drop in Monolithic 3D ICs

The observations from the dynamic rail analysis, and time- and frequency-domain analysis on M3D ICs are summarized as follows:

- Ldi/dt-drop in M3D ICs is less than 2D ICs, showing only 11.3% higher dynamic voltage drop, which includes both IR-drop and Ldi/dt-drop, than the 2D counterparts
- 3D placement of decap cells prevents M3D ICs suffering from sudden Ldi/dt-drop. Cells in an M3D design utilize decaps placed on the same tier (the xy-plain) as well as on the different tier (the z-axis).
- Improved resiliency against AC current noise helps M3D ICs mitigate the voltage drop, which results from higher resistance of M3D PDN due to series connection across tiers. M3D PDN shows 35.9% peak impedance reduction at the worst-case resonance oscillations.

4.3.3 Monolithic 3D IC Power Delivery Network Optimization

The optimization methodologies to improve voltage drop in M3D ICs are summarized as follows:

- M3D PDN optimization should focus on reducing IR-drop of M3D ICs since M3D PDNs show their strength on L$\mathrm{d}i/\mathrm{d}t$-drop
- Top-tier cell repositioning helps current flow the minimum resistive path without being blocked by top-tier cells. It avoids top-tier cell placement on the location that power MIVs are to be placed, showing up to 14.8% static voltage drop reduction.
- Asymmetric top- and bottom-tier M3D PDN in conjunction with top-tier cell repositioning technique reduces static voltage drop further, showing 25.7% higher static voltage drop and 9.0% less dynamic voltage drop compared to its 2D counterpart. The technique utilizes wider width for top-tier metal layers in M3D PDN since higher current flows in the metal layers than those of the bottom tier, whereas narrower bottom-tier power rails to retrieve signal routing resources taken by wider top-tier power rails.

CHAPTER 5

MONOLITHIC 3D ICS FOR DEEP NEURAL NETWORK HARDWARE

5.1 Motivation and Background

DNNs have become ubiquitous in many machine learning applications, from speech recognition [30, 31] and natural language processing [32], to image recognition [33, 34], and computer vision [35]. Large neural network models have proven to be very powerful in all the stated cases, but implementing high-speed (i.e., high-performance), energy-efficient DNN ASIC is still challenging because (1) the required computations consume large amounts of processing time and energy, (2) the memory needed to store the weights are prohibitive, and (3) excessive wire overhead exists due to a large number of connections between neurons, which makes a DNN ASIC a heavily wire-dominated circuit.

Modern DNNs may require $>100M$ parameters [36] for large-scale speech recognition tasks. This is impractical using only on-chip memory due to power density and temperature instability [37], and hence offloading storage to an external DRAM is required. With the introduction of an external DRAM, however, the bottleneck for computation efficiency is now determined by the parameter fetching from DRAM [38]. To mitigate this bottleneck, recent works have compressed the neural network weights 'in architectural perspective' and substantially reduced the amount of computation required to obtain the final output [39, 40, 41, 42], which becomes crucial for efficient DNN ASICs. An alternate method of reducing the complexity caused by the vast requirement of memory for DNNs is in-training quantization of the network parameters [43, 44]. This method, however, is not explored in this work.

With the weight-compressed DNN architecture, M3D stacking technology is adopted to further improve the energy-efficiency and performance 'in physical design perspective'.

5.2 Impact of Monolithic 3D ICs on On-Chip Deep Neural Networks Targeting Speech Recognition

In this work, the impact of M3D stacking technology on power, performance, and area is investigated with speech recognition DNN architectures that exhibit coarse-grain sparsity. M3D ICs reduce the total power consumption more effectively with compute-intensive workloads, compared to memory-intensive workloads. By placing memory blocks evenly on both tiers, M3D ICs reduce the total power consumption up to 22.3%. In addition, owing to the reduced footprint and vertical integration, M3D ICs offer performance improvement over 2D ICs, especially in architecture with complex combinational logics.

5.2.1 Deep Neural Network for Speech Recognition

Topology, the training and classification strategy are presented, which are used for the DNN architectures for speech recognition used in this work. In addition, the coarse-grain sparsification (CGS) is introduced, which effectively reduces area and computation overhead of DNNs are presented.

DNN Topology

Starting from a fully-connected DNN, a Gaussian Mixture Model (GMM) is adopted for acoustic modeling [45]. Since it has been shown that DNNs in conjunction with Hidden Markov Models (HMMs) increase recognition accuracy [30], a HMM is also employed to model the sequence of phonemes. The most likely sequence is determined by the HMM utilizing the Viterbi algorithm for decoding. Then, the CGS methodology presented in [41] is adopted in the DNN architecture to reduce the memory footprint as well as the computation for DNN classification.

As shown in Figure 5.1, the DNN for speech recognition consists of 4 hidden layers with 1,024 neurons per layer. There are 440 input nodes corresponding to 11 frames (5 previous,

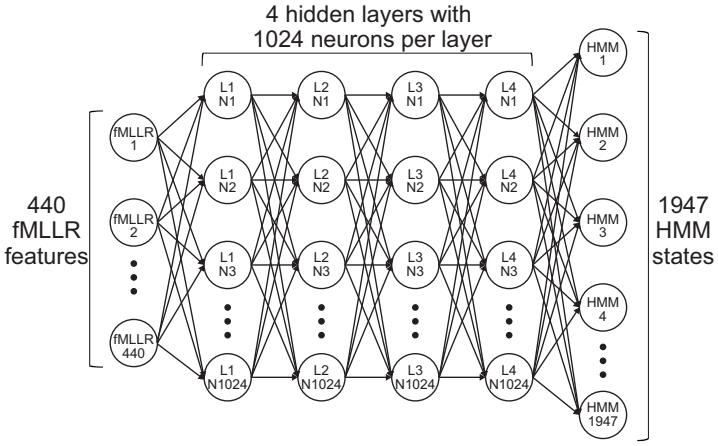

Figure 5.1: Diagram of the DNN for speech recognition

5 future, and 1 current) with 40 feature-space Maximum Likelihood Linear Regression (fMLLR) features per frame. The output layer consists of 1,947 probability estimates, and they are sent to the HMM unit to determine the best sequence of phoneme using the TIMIT database [46]. The Kaldi toolkit [47] is utilized for the transcription of the words and sentences for the particular set of phonemes.

Deep Neural Network Training and Classification

The DNN is trained with the objective function that minimizes the cross-entropy error of the outputs of the network, as described in Equation (5.1).

$$E = -\sum_{i=1}^{N} t_i \cdot ln(y_i), \tag{5.1}$$

where N is the size of the output layer, y_i is the i^{th} output node, and t_i is the i^{th} target value or label. The mini-batch stochastic gradient method [48] is used to update the weights. The weight W_{ij} is updated in the $(k+1)^{th}$ iteration using Equation (5.2).

$$(W_{ij})_{k+1} = (W_{ij})_k + C_{ij}(-lr(\Delta W_{ij})_k + m(\Delta W_{ij})_{k-1}), \tag{5.2}$$

where m is the momentum, lr is the learning rate, and C_{ij} is the binary connection coefficient between two subsequent neural network layers for CGS. In CGS, only the weights that correspond to the location where $C_{ij} = 1$ are updated. The change in weight for each

iteration is the differential of the cost function with respect to the weight value:

$$\Delta W = \frac{\delta E}{\delta W}, \tag{5.3}$$

such that the loss reduces in each iteration. The training procedure is performed on a graphics processing unit (GPU) with 32-bit floating point values.

After training, feed-forward computation is performed for classification, through matrix-vector multiplication of weight matrices and neuron vectors in each layer to obtain the output of the final layer. The Rectified Linear Unit (ReLU) function [33] is used for the non-linear activation function at the end of each hidden layer.

Coarse-Grain Sparsification

To efficiently map sparse weight matrices to memory arrays, CGS methodology [41] is employed. In CGS, connections between two consecutive layers in a DNN are compressed in a block-wise manner. An example of block-wise weight compression is demonstrated in Figure 5.2. For a given block size of 16×16, it reduces a 1,024×1,024 weight matrix to 64×64 weight blocks. With a compression ratio of 87.5%, only eight weight blocks (= 12.5%) remain non-zero for each block row, thus allowing for efficient compression of the entire weight matrix with minimal index.

CGS, when compared to recent neural network compression algorithms such as in [49, 50], offers simpler hardware implementation through CGS multiplexers and multiplier-accumulators (MACs). In [49], a complex sparse matrix vector multiplication module is required. On the other hand, the methodology in [50] offers to reduce the order of computations needed for a matrix of size n to $\mathcal{O}(n \log n)$ and reduce the space required to store the matrix to $\mathcal{O}(n)$. However, there is considerable loss in accuracy when the size of the matrix increases, and hardware for computing FFT and inverse fast fourier transform (IFFT) is required. The issue of matrix size is resolved in [51] using block-circulant matrices, but the advantage of using FFT and IFFT to compute matrix vector multiplications is lost if the size of the blocks reduce significantly. This restriction is not present if CGS is used.

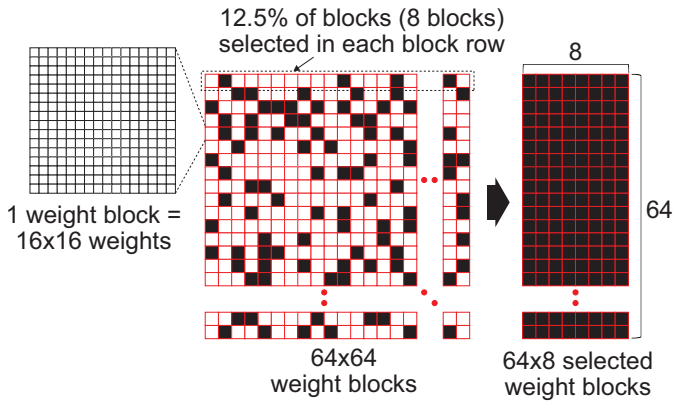

Figure 5.2: An example of block-wise weight compression in CGS. 1,024×1,024 weight matrix is divided into 64×64 weight blocks with each weight block having 16×16 weights (i.e. block size of 16×16). 87.5% of weight blocks are dropped using CGS. The remaining 12.5% weight blocks are stored in memory.

Table 5.1: Key parameters of the two CGS-based DNN architectures used in this work: DNN CGS-16 and DNN CGS-64

parameter	DNN CGS-16	DNN CGS-64
block size	16×16	64×64
compression rate	87.5%	87.5%
phoneme error rate	19.8%	19.9%

GPU-accelerated DNN computations can also benefit from CGS. With CGS, along with the testing inference, training complexity can also be reduced due to the sparse nature of the weight matrices. The structured sparseness allows for writing customized GPU kernels that only need to operate on the non-zero elements, significantly speeding up training and reducing GPU power consumption as shown in [52].

In order to study the impact of M3D ICs on the power, performance, and area of different DNN architectures, the block sizes are swept for the compression ratio of 87.5%, and the two DNN architectures that have the two lowest phoneme error rates (PER) for the TIMIT dataset are selected for hardware implementation. The two architectures chosen are the DNN with 16×16 block size (DNN CGS-16) and the DNN with 64×64 block size (DNN CGS-64), as shown in Table 5.1.

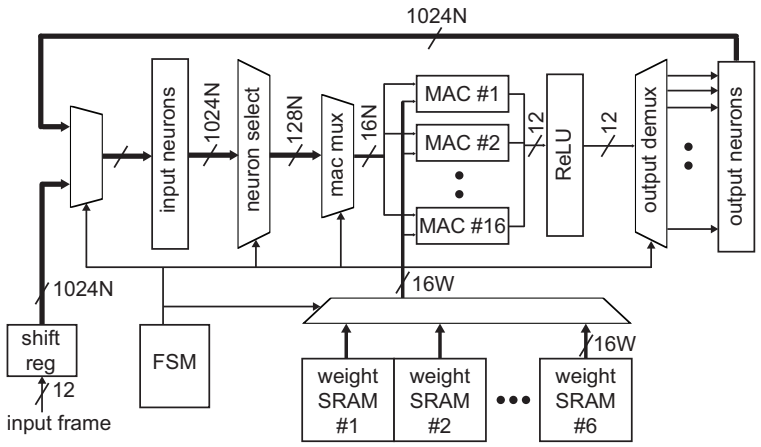

Figure 5.3: Block diagram of the CGS-based DNN architecture for speech recognition

5.2.2 Deep Neural Network Architecture Description

The block diagram of the CGS-based DNN architecture is shown in Figure 5.3. The DNN operates on one layer at a time and consists of 16 MAC units that operate in parallel. The weights of the network are stored in the SRAM banks, while the input and output neurons are stored in registers. The finite state machine (FSM) coordinates the data flow such as layer control and computational resource allocation (i.e., MAC units).

Since the target compression ratio of the architectures is 87.5%, the neuron select unit chooses 128 neurons (12.5%) among 1,024 input neurons that proceed to the MAC units. This selection-based computation eliminates unnecessary MAC operations (i.e., MAC operation of neurons corresponding to zero weights in CGS-based weight matrix). The neuron select unit is controlled by the binary connection coefficients discussed in Section 5.2.1, and the coefficients are stored in the dedicated register file in the FSM unit.

The size of the register file is determined by the block size used in the DNN architecture. For example, for each hidden layer, eight weight blocks per each row of 64×64 weight blocks are selected for MAC operation in the DNN CGS-16 architecture (Figure 5.2). Thus, eight multiplexers are required in the neuron select unit, and each multiplexer selects one weight block among 64 in a block row, so that each multiplexer requires six selection bits ($=\log_2 64$). Since there are 64 total block rows in the architecture, the total number of bits

to obtain 64×8 selected weight block for a hidden layer is 3,072 bits (= eight multiplexers × 6 selection bits × 64 block rows). Although the DNN has four hidden layers, the number of coefficients for the last hidden layer should be doubled because the number of neurons in the output layer (1,947 HMM states) is almost 2× of other layers. Therefore, the size of the coefficient register file in the DNN CGS-16 is 15,360 bits (= 3,072 bits × 5 effective layers). This value is calculated in the same way for the DNN CGS-64 architecture, resulting in 640 bits in total.

On-chip SRAM arrays store the compressed weight parameters in six banks for the four hidden layers and the output layer (∼2× parameters). The size of the SRAM bank is determined by the number of MAC units in the architecture. Since the DNN architectures operate 16 units in parallel, the row size of each SRAM bank is 128 bits (= 16 MAC units × 8-bit weight precision). Since 8,192 rows are assumed for each SRAM bank, the total size of the six SRAM banks in the DNN is 6Mb (= 6 banks × 128 bits × 8,192 rows). This compact memory size with the CGS methodology enables the DNN to store the compressed weight parameters on chip.

5.2.3 Impact of Monolithic 3D ICs on Energy-Efficiency of Deep Neural Network Hardware

To analyze the advantage of M3D stacking technology on energy-efficiency of different DNN architectures, two DNN architectures (DNN CGS-16 and CGS-64) are implemented using TSMC® $28nm$ HPM technology with a target clock frequency of $400MHz$. The footprint of 2D ICs are set by targeting cell density of 65%. M3D ICs are implemented using **shrunk-2D design flow** presented in Section 2.2.1. The impact of memory tier partitioning scheme is examined by comparing two memory floorplan schemes for M3D ICs, one with memory blocks on both tiers (i.e., M3D-both), and the other with memory blocks on a single tier only (i.e., M3D-one). In the M3D-both designs, memory blocks are evenly split on the top and bottom tiers using similar floorplan for both tiers. On the other hand, in the M3D-one designs, all standard cells are placed on one tier, and only memory blocks

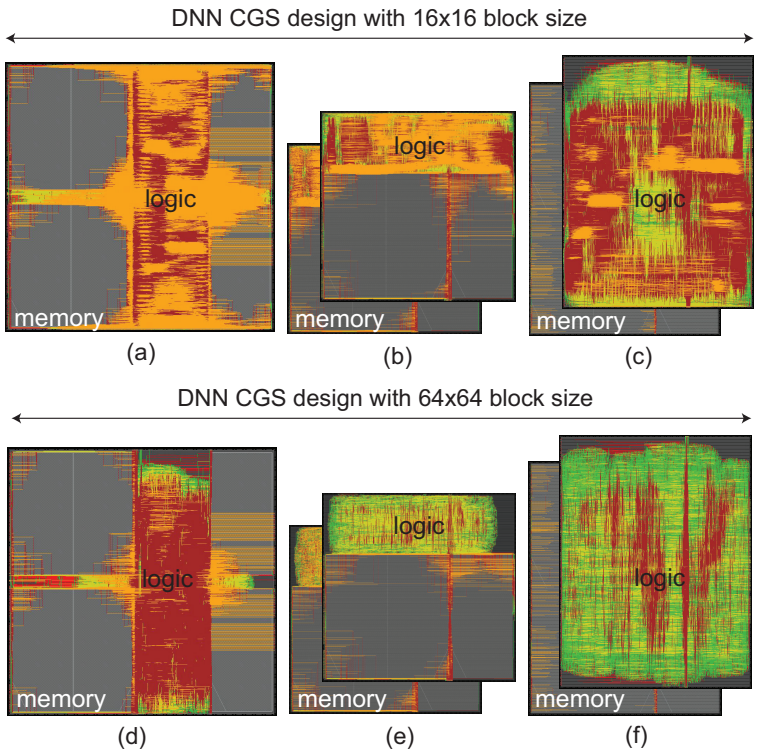

Figure 5.4: GDS layouts of the implemented DNN CGS-16 and CGS-64 architectures at $400MHz$ target clock frequency. DNN CGS-16 (a) 2D IC, (b) M3D-both, (c) M3D-one, DNN CGS-64 (d) 2D IC, (e) M3D-both, (f) M3D-one.

exist on the other tier. Figure 5.4 shows the GDS layouts of the implemented 2D and M3D ICs.

Area, Wire-length, and Capacitance Comparisons

Iso-performance comparison of several key metrics of the 2D and M3D ICs is presented in Table 5.2. The M3D-both designs achieve 50.1% footprint reduction compared with the 2D ICs, whereas the M3D-one designs obtain only 33.9% reduction. This difference is attributed to the large memory area compared with logic in the DNN CGS-16 2D IC. These large memory blocks, if placed in the same tier, cause the footprint to increase significantly.

The wire-length saving reaches 29.9% and 33.7% in CGS-16 and CGS-64, respectively, with the M3D-both designs. This significant wire-length saving comes from the

Table 5.2: Iso-performance ($400MHz$) design metric comparison of 2D and M3D ICs of DNN CGS-16 and CGS-64 architectures. All percentage values show the reduction from their 2D counterparts.

parameter	2D	M3D-both		M3D-one	
DNN CGS-16					
footprint (μm)	1,411×1,411	1,010×984	(-50.1%)	996×1,322	(-33.9%)
wire-length (m)	12.089	8.469	(-29.9%)	12.225	(1.1%)
cell count	298,309	262,084	(-12.1%)	290,692	(-2.6%)
cell area (mm^2)	0.505	0.431	(-14.6%)	0.511	(1.1%)
mem area (mm^2)	1.287	1.287	(0.0%)	1.287	(0.0%)
MIV count	-	77,536		1,776	
pin cap (pF)	943.3	788.0	(-16.5%)	1,004.1	(6.4%)
wire cap (pF)	2,216.8	1,440.8	(-35.0%)	2,087.4	(-5.8%)
total cap (pF)	3,160.1	2,228.7	(-29.5%)	3,091.6	(-2.2%)
DNN CGS-64					
footprint (μm)	1,411×1,411	1,010×984	(-50.1%)	996×,1322	(-33.9%)
wire-length (m)	5.631	3.734	(-33.7%)	7.134	(26.7%)
cell count	163,361	149,921	(-8.2%)	174,292	(6.7%)
cell area (mm^2)	0.314	0.269	(-14.3%)	0.328	(4.7%)
mem area (mm^2)	1.287	1.287	(0.0%)	1.287	(0.0%)
MIV count	-	48,636		1,776	
pin cap (pF)	520.8	390.8	(-25.0%)	553.5	(6.3%)
wire cap (pF)	920.1	573.7	(-37.7%)	1,110.5	(20.7%)
total cap (pF)	1,440.9	964.4	(-33.1%)	1,664.0	(15.5%)

50% smaller footprint and shorter distance among cells in M3D ICs. The M3D-both design for CGS-16 architecture achieves 12.1% cell count reduction, which leads to 14.6% total cell area saving. This saving mainly comes from fewer buffers and smaller gates needed to close timing in M3D ICs compared with the 2D counterparts. The savings in CGS-64 architecture are 8.2% and 14.3% for the cell count and area, respectively.

$77K$ MIVs are utilized in the CGS-16 architecture, while $48K$ MIVs are used in CGS-64. This is mainly because CGS-16 design is more complex than CGS-64 (to be further discussed in Section 5.2.5), so that the tier partitioning cutline cuts through more inter-tier connections in CGS-16. In the M3D-one design, logic and memory are separated into different tiers. This logic-memory connectivity is not high in the DNN architecture (= $1.7K$).

In the CGS-16 architecture, the 16.5% pin capacitance saving is from cell area reduc-

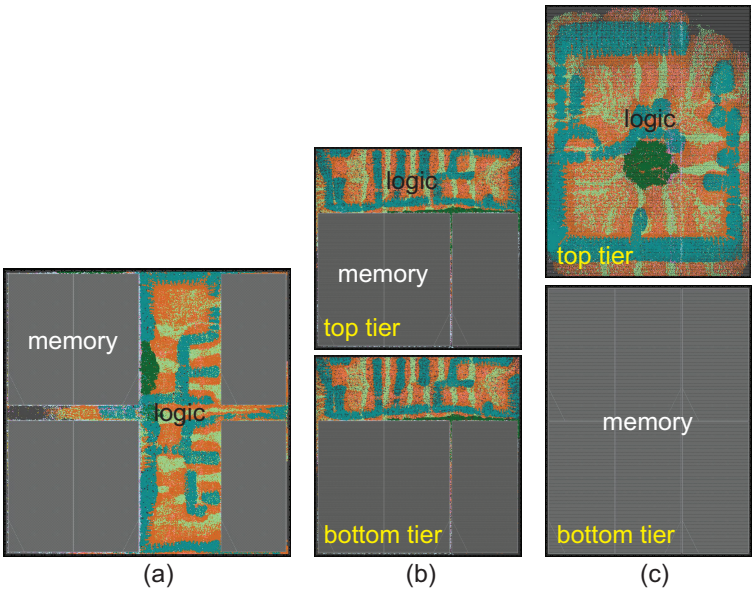

Figure 5.5: Cell placement of the modules in CGS-16 architecture. (a) 2D, (b) M3D-both, (c) M3D-one. Each module is highlighted with different colors.

tion, while the 35.0% wire capacitance saving is from wire-length reduction. By comparing the raw data, the DNN architecture is wire-dominated. The pin and wire capacitance saving reaches 25.0% and 37.7% in CGS-64.

To better understand why M3D-one gives significantly worse results than M3D-both, a placement comparison among 2D, M3D-both, and M3D-one designs is shown in Figure 5.5. In the M3D-both design shown in Figure 5.5 (b), the logic cells related to memory blocks in the top tier are placed in the same tier as the memory and densely packed to reduce wire-length effectively. This is the same for the bottom tier in the M3D-both design. On the other hand, logic gates are rather spread out across the top tier in the M3D-one design shown in Figure 5.5 (c). This results in 1.1% increase in wire-length for CGS-16 and 26.7% increase in wire-length for CGS-64 compared with the 2D counterparts. This highlights the importance of footprint management and tier partitioning in the presence of large memory modules in DNN architectures.

Table 5.3: Iso-performance ($400MHz$) power metric comparison of two architectures (CGS-16 vs. CGS-64) using two workloads (classification vs. pseudo-training). All percentage values show the reduction from their 2D counterparts.

workload	power breakdown	2D	M3D-both		M3D-one	
DNN CGS-16						
classification	internal power (mW)	91.3	76.7	(-16.0%)	90.3	(-1.1%)
	net switching power (mW)	48.6	31.6	(-35.0%)	46.5	(-4.3%)
	leakage power (mW)	1.3	1.2	(-6.6%)	1.3	(0.5%)
	total power (mW)	**141.1**	**109.6**	**(-22.3%)**	138.0	(-2.2%)
pseudo-training	internal power (mW)	150.4	142.8	(-5.1%)	148.3	(-1.4%)
	net switching power (mW)	68.4	57.1	(-16.6%)	65.6	(-4.2%)
	leakage power (mW)	1.3	1.2	(-6.8%)	1.3	(0.7%)
	total power (mW)	**220.0**	**201.0**	**(-8.6%)**	215.0	(-2.3%)
DNN CGS-64						
classification	internal power (mW)	86.8	76.1	(-12.3%)	84.9	(-2.2%)
	net switching power (mW)	41.2	30.2	(-26.7%)	42.8	(3.9%)
	leakage power (mW)	1.1	1.1	(-4.7%)	1.1	(1.5%)
	total power (mW)	**129.1**	**107.3**	**(-16.9%)**	128.8	(-0.2%)
pseudo-training	internal power (mW)	129.2	120.0	(-7.2%)	128.5	(-0.5%)
	net switching power (mW)	46.0	36.3	(-21.2%)	50.3	(9.3%)
	leakage power (mW)	1.1	1.1	(-4.6%)	1.1	(1.4%)
	total power (mW)	**176.3**	**157.4**	**(-10.7%)**	179.9	(2.0%)

Power Comparisons

Table 5.3 presents the iso-performance power comparison between 2D and M3D ICs of CGS-based DNNs. internal, net switching, and leakage power breakdown is reported for each design. The sign-off power calculations are conducted using two speech recognition workloads: classification and pseudo-training (more details provided in Section 5.2.5).

During classification, CGS-16 consumes $141.1mW$, while CGS-64 consumes $129.1mW$. This confirms that CGS-16 consumes more power to handle more complex weight selection process (to be further discussed in Section 5.2.5). A similar trend is observed during pseudo-training.

Pseudo-training, as expected, causes more switching in the circuits, and thus more power consumption compared with classification for both CGS-16 and CGS-64 architectures.

Next, the power consumption of 2D and M3D ICs are compared. The resulting foot-

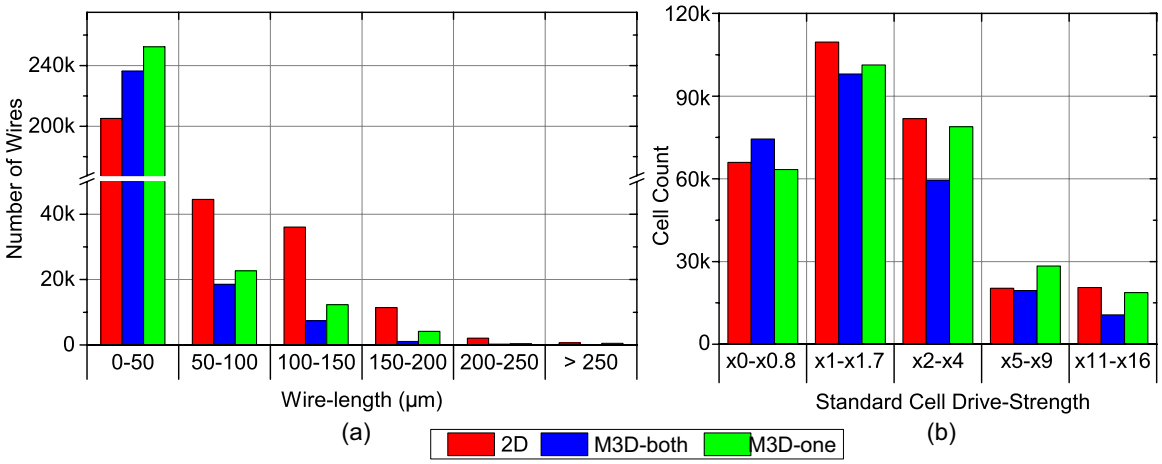

Figure 5.6: (a) Wire-length and (b) cell drive-strength distribution of DNN CGS-16 2D, M3D-both, and M3D-one

print of M3D-both designs is reduced by half, thereby reducing the wire-length between the cells. Figure 5.6 (a) shows the wire-length distribution of the 2D and M3D ICs of CGS-16 architecture. The histogram clearly shows that M3D ICs contain more number of short wires and fewer long wires compared with 2D IC. The effect of wire-length saving translates to the reduction of wire capacitance C_{wire} in Equation (2.8), therefore the saving of the third term of the equation. Figure 5.6 (b) presents the distribution of standard cells with different ranges of cell drive-strength. M3D-both design uses more number of low drive-strength cells (i.e., $\times 0 \sim \times 0.8$) and fewer high drive-strength cells (i.e., $\times 1 \sim \times 16$). Since low drive-strength cells utilize smaller transistors, the short circuit current of the transistors and C_{pin} are lower, which reduces both the first and second term in Equation (2.8).

5.2.4 Impact of Monolithic 3D ICs on Performance of Deep Neural Network Hardware

In this section, the impact of M3D stacking technology on the performance of CGS-16 and CGS-64 architectures is investigated by pushing the target clock frequency of 2D and M3D ICs to their maximum clock frequency. 2D and M3D ICs are implemented with TSMC® $28nm$ HPM technology sweeping the target frequency from $400MHz$ in $25MHz$ increments. The floorplans of the 2D and M3D ICs are same as the ones used in Section 5.2.3.

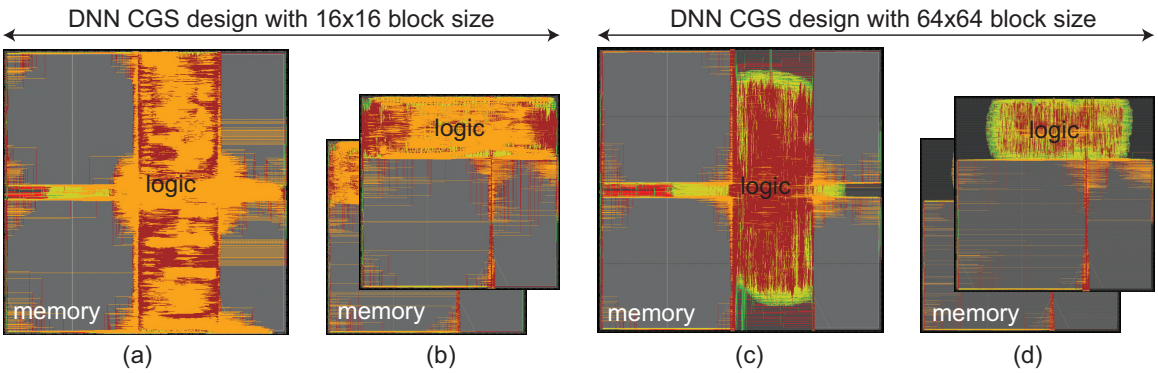

Figure 5.7: GDS layouts of 2D and M3D ICs of DNN CGS-16 and CGS-64 architectures at the maximum target frequencies. (a) 2D IC at 550MHz, (b) M3D IC at 575MHz of DNN CGS-16 architecture, (c) 2D IC at 600MHz, (d) M3D IC at 625MHz of DNN CGS-64 architecture.

Table 5.4: Maximum performance comparison of 2D and M3D ICs of DNN CGS-16 and CGS-64 architectures

parameter		DNN CGS-16	DNN CGS-64
2D	target clk freq (MHz)	550	600
	WNS (ns)	-0.056	0.002
	effective clk freq (MHz)	**534**	**601**
M3D	target clk freq (MHz)	575	625
	WNS (ns)	-0.024	-0.046
	effective clk freq (MHz)	**567**	**608**
Δ% effective clk freq		**6.2%**	**1.2%**

As the M3D-both designs show better design quality compared to the M3D-one designs, memory blocks are placed on both tiers in the M3D ICs for this experiment as shown in Figure 5.7.

The maximum performance comparison between the 2D and M3D ICs of CGS-16 and CGS-64 architectures is presented in Table 5.4. The table shows the target clock frequency used to place-and-route the designs, the resulting WNS from static timing analysis, and the effective clock frequency, which is the maximum achievable clock frequency that the designs are able to operate at without timing violation.

Comparing only the 2D ICs of CGS-16 and CGS-64 architectures, the effective clock frequency of the CGS-16 2D IC is 11.1% less than the CGS-64 2D IC. As the critical path of the CGS-16 2D IC starts from weight SRAM to MAC unit through weight selection logic,

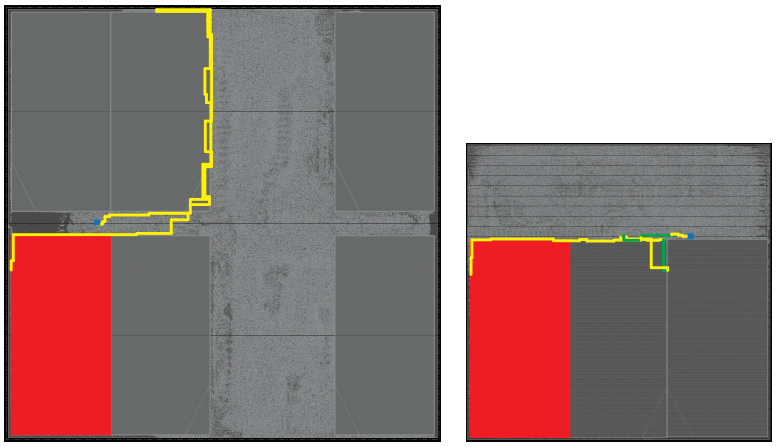

Figure 5.8: Worst timing path comparison of 2D and M3D ICs of DNN CGS-16 architecture. (a) The worst timing path of 2D IC at its maximum target clock frequency, $550MHz$. (b) The same timing path in M3D IC. Cells in the top tier are projected into the bottom tier for M3D IC, and red boxes (i.e., weight SRAM) indicate the start point, whereas blue boxes (i.e., flip-flops in MAC unit) represent the end point of the timing path. Yellow lines show the wires in 2D and the bottom tier of M3D IC, whereas green lines are the top tier wires in M3D IC.

the lower effective clock frequency of the CGS-16 2D IC is attributed to its more complex weight selection logic as shown in a higher design density in Figure 5.7 (a) compared to Figure 5.7 (c).

Next, the maximum performance of the 2D and M3D ICs is compared. The M3D ICs shows 6.2% and 1.2% performance improvement over 2D counterparts in CGS-16 and CGS-64 architectures, respectively. To analyze this trend, the worst timing path comparison of the 2D and M3D ICs is conducted. Figure 5.8 compares the same timing path (i.e., the worst timing path of the 2D IC) in the 2D and M3D CGS-16 designs at the maximum target clock frequency of the 2D IC, and Table 5.5 presents key metrics of the timing path.

The wire-length of the worst timing path of the 2D IC is 53.6% longer than the same timing path in the M3D IC. This is attributed to the reduced footprint and the inter-tier connections of the M3D IC, which results in shorter distance among cells along the timing path. The M3D IC offers 24.6% cell count saving as well as 21.3% average cell drive-strength reduction, thereby reducing cell area by 63.1% of the timing path. This is because

Table 5.5: Key parameter comparison of the worst timing path in Figure 5.8 of the 2D and M3D ICs of DNN CGS-16 architecture

parameter	2D	M3D	
wire-length (μm)	3,208	1,488	(-53.6%)
cell count	65	49	(-24.6%)
avg. cell drv-str	8.9	7.0	(-21.3%)
cell area	180.1	66.4	(-63.1%)
MIV count	-	6	
wire cap (fF)	500	242	(-51.6%)
pin cap (fF)	486	312	(-35.8%)
resistance ($k\Omega$)	14.5	9.3	(-35.9%)
delay (ns)	**2.344**	**2.088**	**(-10.9%)**

the fewer and smaller buffers are needed to drive the reduced wire-load, which is a result of the wire-length reduction.

Compared to the 2D IC, the wire and pin capacitance of the timing path in the M3D IC are reduced by 51.6% and 35.8%, respectively. The wire capacitance reduction mainly comes from the wire-length reduction of the timing path, whereas the pin capacitance saving results from the cell count and cell drive-strength reduction. In addition, the M3D IC achieves 35.9% resistance reduction in the timing path. The resistance saving is also attributed to the wire-length saving along the timing path.

Due to the capacitance and resistance saving of the worst timing path, the delay of the timing path is reduced by 10.9% in the M3D IC, thereby offering rooms to improve the performance.

In order to understand the impact of the above observations to the overall timing paths of the 2D and M3D ICs, the slack distribution of all timing paths of the 2D and M3D CGS-16 designs is reported in Figure 5.9. While 18 timing paths of the 2D IC violate the timing constraints, the M3D IC successfully closes timing without any violation. In addition, there are more timing paths with high positive slack in the M3D IC, which indicates that timing is easily closed in the M3D IC due to the reduced delay of the timing paths.

The difference in the performance improvement of the M3D ICs of CGS-16 and CGS-64 architecture is also attributed to the complex weight selection logic in CGS-16 and will

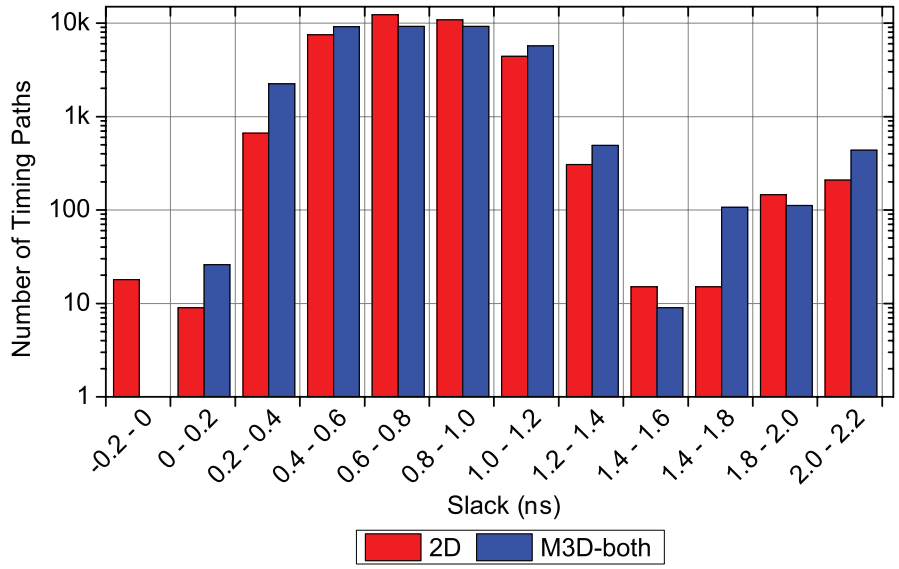

Figure 5.9: Slack distribution comparison between 2D and M3D ICs of DNN CGS-16 architecture at the maximum clock frequency of the M3D IC.

be discussed in detail in Section 5.2.5.

5.2.5 Architectural Impact Discussions

CGS-16 and CGS-64 Architecture Comparisons

Table 5.3 shows that the total power reduction of M3D ICs is higher in DNN CGS-16 architecture than CGS-64. Furthermore, more performance improvement with M3D ICs is achieved in DNN CGS-16 architecture as shown in Table 5.4. These differences are caused by the granularity of weight selection methodology (i.e., CGS algorithm). The 1,024×1,024 weight matrix is divided into 256 (= 16×16) weight blocks in CGS-64 architecture. This count becomes 4,096 (= 64×64) weight blocks in CGS-16. The implication in DNN architecture is that CGS-16 requires a more complex neuron selection unit than CGS-64. Figure 5.10 shows the comparison of standard cell area of each module in CGS-16 and CGS-64 architectures. Both sequential (dashed box) and combinational logic (non-dashed box) portion in each module are shown. The neuron selection unit in CGS-16 architecture (shown in purple) occupies more area than that in CGS-64 architecture.

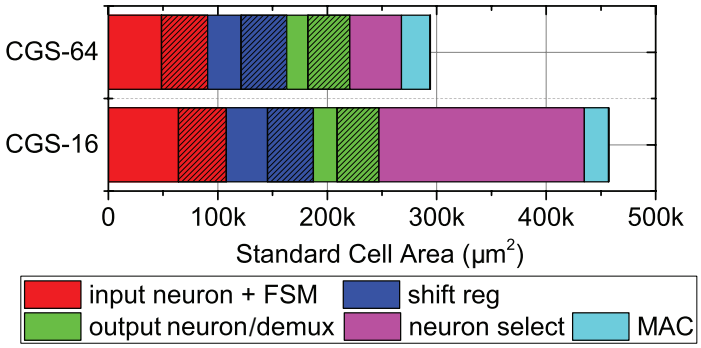

Figure 5.10: Standard cell area breakdown of 2D CGS-16 and CGS-64 architectures. Non-dashed and dashed boxes respectively indicates combinational and sequential elements. Only five largest modules are shown.

As discussed in Section 5.2.3, M3D ICs benefit not only from wire-length reduction but also from standard cell area saving. The number of storage elements (i.e., sequential logic and memory blocks) used in 2D and M3D ICs remain the same. Thus, the only possible power reduction coming from storage elements is their drive-strength reduction. This does not show a huge impact considering the small portion of sequential elements in the DNN architectures (16.1% on average). On the other hand, combinational logic can be optimized in various ways, such as logic reconstructing and buffer reduction. Therefore, the DNN M3D ICs benefit more from combinational logic gates than sequential elements.

Figure 5.11 shows the breakdown of total power consumption into combinational, register, clock, and memory portions. Combinational power reduction is the dominant factor in total power saving of M3D ICs in both CGS-16 and CGS-64 architectures and in both classification and pseudo-training workloads. The saving in other parts including register, clock, and memory power largely remain small. In addition, the neuron selection unit in CGS-16 architecture consists of a larger number of combinational logic gates than CGS-64. Thus, its M3D ICs have more room for power optimization, resulting in a larger combinational power saving.

The larger neuron selection logic in CGS-16 architecture also offers more opportunity to improve the performance of M3D ICs. While 2D ICs suffer long timing path due to the complex neuron selection logic, M3D ICs effectively reduce the wire-length, providing

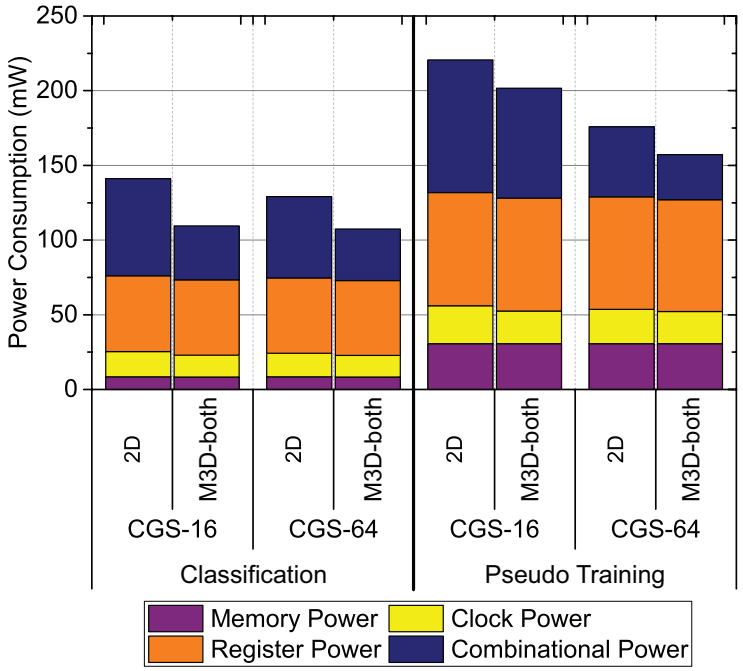

Figure 5.11: Power breakdown under two DNN architectures (CGS-16 and CGS-64), two workloads (classification and pseudo-training), and two designs (2D and M3D ICs)

buffer count/size reduction along the worst timing path. This reduces the capacitance and resistance of timing paths, thereby offering shorter delay and larger performance improvement.

Figure 5.12 compares the total wire-length and standard cell count along the selected 486 timing paths, which are from weight SRAMs to registers of MAC units through neuron selection logic, in the 2D/M3D CGS-16/CGS-64 designs at the maximum frequency of the 2D ICs. Comparing only the 2D ICs, the CGS-16 2D IC clearly utilizes longer wire-length as well as more standard cells as the neuron selection logic is more complex. As the CGS-16 M3D IC has more combinational logics to optimize with the reduced footprint, it offers more cell count and wire-length reduction compared to the CGS-64 M3D IC, providing more rooms for performance improvement in higher clock frequency.

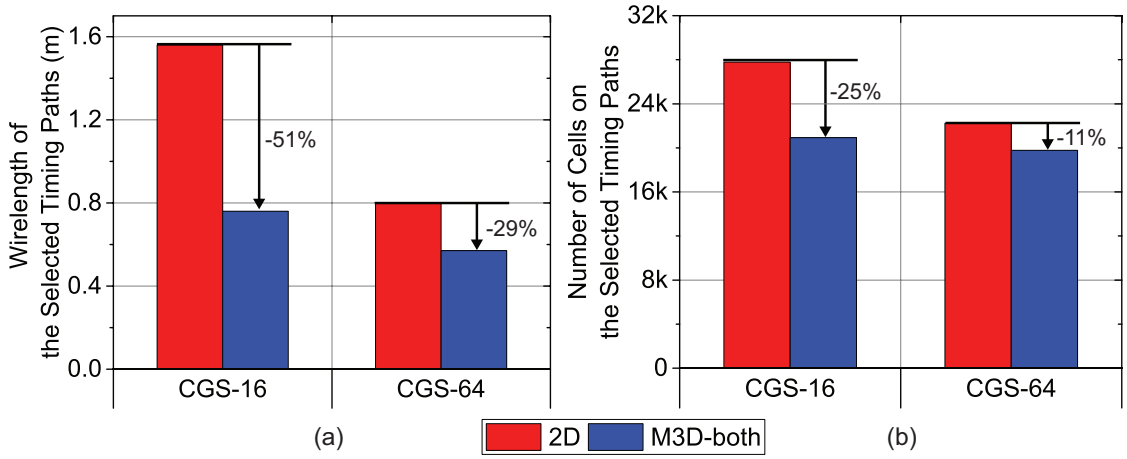

Figure 5.12: Comparison of (a) the wire-length and (b) cell count of the timing paths from weight SRAMs to registers in MAC units through neuron selection logic in 2D and M3D-both of DNN CGS-16 and CGS-64 architecture

Impact of Workloads

In order to investigate the impact of different DNN workloads on M3D power saving, two main types of speech recognition DNN workloads are analyzed: feed-forward classification and training. Real-world test vectors are used for feed-forward classification. However, since the current architecture does not supports online training to avoid computational overhead of finding gradients in DNN training, customized test vectors are created for 'pseudo-training'. Online training on DNN consists of feed-forward computation and backward computation. In order to mimic the online training on the current architecture, there are two phases in the pseudo-training test vectors as shown in Figure 5.13. In the first phase, the DNN performs feed-forward classification, which represents feed-forward computation during training. In the second phase, the DNN conducts feed-forward classification and writes the weights to memory blocks, which represents backward computation and weight update. These two phases mimic the behavior of logic computation and weight update during training.

Table 5.3 shows that while M3D-both shows 22.3% (CGS-16) and 16.9% (CGS-64) total power reduction in feed-forward classification workload, the power saving of pseudo-

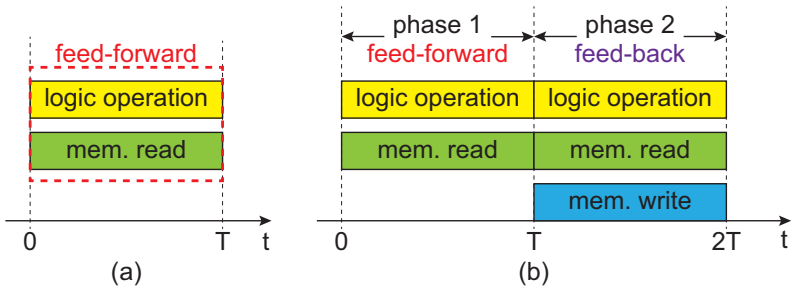

Figure 5.13: Comparison of the operations in (a) the feed-forward classification and (b) pseudo-training

training workload is only 8.6% (CGS-16) and 10.7% (CGS-64). This difference stems from different switching patterns of combinational logic and storage elements in the DNN architecture. The DNN mainly uses combinational logic gates to compute the values of neuron outputs and access memory for read operations only during feed-forward classification. Thus, this workload is classified as a compute-intensive kernel. On the other hand, memory operations are heavily used during pseudo-training since the DNN architecture needs to read and write weights. This becomes a memory-intensive kernel. Therefore, switching activity in memory blocks is much higher during pseudo-training while that of combinational logic remains largely similar. This explains larger power consumption during pseudo-training workload: 220.0mW vs. 141.1mW for CGS-16, and 176.3mW vs. 129.1mW for CGS-64 as shown in Table 5.3.

As shown in Figure 5.11, memory power and register power occupy a large portion of the total power during pseudo-training. This means that the combinational logic power saving becomes a smaller portion of the total power saving during training. The opposite is true for classification, where memory and register power are less dominant. In this case, the reduction in combinational power saving becomes more prominent in the total power saving.

5.3 Summary

The lessons learned from this study and the design guidelines to maximize the power benefits of M3D ICs targeting DNN architectures are as follows:

- M3D stacking technology effectively reduces the total power consumption of DNN architectures by reducing wire-length as well as standard cell area, showing its efficacy on saving power consumption of wire-dominated DNN circuits
- M3D ICs enhance the performance of DNN designs. This is mainly attributed to the reduced capacitance and resistance of timing paths in the designs, which comes from both wire-length and buffer count/size reduction.
- If memory blocks occupy more than half area of a DNN design, partitioning the memory blocks onto two tiers (i.e., M3D-both designs), instead of placing them on one tier (i.e., M3D-one designs), helps maximize the total power saving of the M3D IC. It is because M3D-both designs achieve smaller footprint in that case, which makes cell placement denser, and hence, reduces more wire-length.
- M3D ICs show larger power savings with smaller CGS block sizes, which consists of more combinational logics, in speech recognition DNNs. This enables the choice of selecting smaller block sizes for CGS in hardware implementations, which was earlier overlooked due to larger power overhead in 2D ICs.
- DNNs with smaller CGS block sizes also benefit more on their performance from M3D ICs, effectively reducing the overhead of more complex weight selection logic
- In the DNN used in this work, it is combinational logic power reduction, not the commonly believed memory-related power reduction, that dominates the overall power saving of M3D ICs. Moreover, compute-intensive classification workload gives more power saving than memory-intensive training workload with M3D ICs. Such a claim cannot become a general statement, and other DNN architectures may prove to be the opposite.

CHAPTER 6

CONCLUSIONS

As device scaling in advanced technology nodes is slowly saturating due to low volume and yields, M3D stacking technology has come into the spotlight as an alternative for continuing Moore's low, showing its strength in reducing power consumption and enhancing performance utilizing short vertical connections among tiers instead of using long wires on the xy-plane. However, There are numerous challenges on M3D stacking technology mainly because of its technological prematurity, sequential fabrication process, and high vertical integration density.

The key design and tool challenges on M3D ICs are summarized as follows: (1) power and performance optimization in advanced technology nodes considering device transition from planar MOSFETs to 3D FinFETs, (2) new M3D IC design flow with micro-architecture-aware partitioning while optimizing multiple tiers simultaneously, (3) optimizations on the PDNs of M3D ICs to resolve higher power density due to multiple layers of transistors.

In this dissertation, in-depth analysis on the above challenges are performed, and physical design optimization methodologies as well as CAD tool solutions are presented in the following projects.

- In-depth analysis and physical design optimization methodologies on performance and power benefits of M3D ICs in advanced technology nodes
- Studies on power benefit trends across technology nodes and a new M3D IC design flow to maximize the benefit
- System-level PDN analysis and optimization for M3D ICs

The impact of M3D ICs on DNN hardware is also presented in this dissertation although it

is not categorized as one of the challenges.

In the first project, comprehensive studies on the factors that impact the performance benefit of M3D ICs are performed, and methodologies to further improve the performance are presented. The optimization methodologies help raise the maximum achievable clock frequencies of M3D ICs and are supported with an in-depth analysis. In addition, the impact of M3D stacking technology on the power efficiency in $7nm$ FinFET based ICs is examined. A predictive $7nm$ PDK and a corresponding cell libraries are developed using commercial-grade tools that accurately model dimensional and material properties accounting for device behavior, cell-level and interconnect parasitics. M3D ICs are implemented and analyzed using the developed $7nm$ PDK and cell libraries. To improve the benefit of M3D ICs further, design space exploration of tier partitioning scheme in M3D IC design flow is performed, and guidelines for bin size selection and prioritized clock tree partitioning methodology are presented.

In the second project, a comprehensive study is performed investigating the power impact of M3D ICs using a commercial in-order 32 bit application processor as the benchmark, implemented on foundry $28nm$, foundry 14/16nm and a predictive $7nm$ technology nodes. M3D ICs provide maximum power savings at the $28nm$ technology node. The benefits improve at higher clock frequencies with the reduction of standard cell area in addition to wire-length savings. Based on the findings, a new M3D IC design flow, called cascade-2D design flow, is developed to implement M3D ICs using 2D commercial tools. Cascade-2D design flow utilizes a design-aware partitioning scheme where functional modules with very large number of connections are partitioned into separate tiers. One of the main advantages of this flow is that it is extremely flexible and is partition-scheme agnostic, making it an ideal methodology to evaluate different M3D partitioning algorithms. The MIVs are modeled as sets of anchor cells and dummy wires, which enable us to implement and optimize both top and bottom tiers simultaneously in a 2D IC. Cascade-2D design flow reduces standard cell area effectively, resulting in significantly better power savings than

existing M3D IC design flows. Additionally, by leveraging smaller standard cells, M3D ICs can save die area which directly translates to reduced costs.

In third, an in-depth study and optimization methodologies for PDNs in M3D ICs are presented. A system-level PDN of M3D ICs is built, and comprehensive studies including static, dynamic rail analysis as well as frequency- and time-domain analysis are performed. Although M3D PDNs suffer from high IR-drop due to additional metal layers, irregular placement of power MIVs, and fewer C4 bumps, they reduce L$\mathrm{d}i/\mathrm{d}t$-drop from 3D placement of decap cells. Additionally, higher resistance of M3D PDN due to its series resistive path across tiers improves the resiliency against AC noise showing peak impedance reduction at first-order resonance frequency. From the observations, two optimization methodologies, top-tier cell repositioning and asymmetric top- and bottom-tier PDNs, are presented, which efficiently improve the power supply integrity of M3D PDNs.

For DNN M3D ICs, the impact of M3D stacking technology on power, performance, and area is examined with speech recognition DNN architectures that exhibit coarse-grain sparsity. M3D ICs reduce the total power consumption more effectively with compute-intensive workloads, compared to memory-intensive workloads. By placing memory blocks evenly on both tiers, M3D ICs reduce the total power consumption significantly. In addition, owing to the reduced footprint and vertical integration, M3D ICs offer performance improvement over 2D ICs, especially in architecture with complex combinational logics. This convincingly demonstrates the low power and high performance benefits of M3D ICs on DNN hardware and offers architectural guidelines to maximize the benefits.

This dissertation demonstrates the potential of M3D ICs and paves the way for more research to combat manufacturing, thermal, process variation and EDA tool challenges associated with M3D stacking technology.

REFERENCES

[1] M. Okada, I. Sugaya, H. Mitsuishi, *et al.*, "High-Precision Wafer-Level Cu-Cu Bonding for 3DICs," in *Proc. Int. Electron Devices Meeting*, 2014.

[2] J. Cong, G. Luo, J. Wei, *et al.*, "Thermal-Aware 3D IC Placement Via Transformation," in *Proc. Asia and South Pacific Design Automation Conf.*, 2007.

[3] G. Katti, M. Stucchi, J. V. Olmen, *et al.*, "Through-Silicon-Via Capacitance Reduction Technique to Benefit 3-D IC Performance," *IEEE Electron Device Letters*, vol. 31, no. 6, pp. 549–551, 2010.

[4] K. Chang, S. Sinha, B. Cline, *et al.*, "Match-Making for Monolithic 3D IC: Finding the Right Technology Node," in *Proc. Design Automation Conf.*, 2016.

[5] S. A. Panth, K. Samadi, Y. Du, *et al.*, "Design and CAD Methodologies for Low Power Gate-level Monolithic 3D ICs," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2014.

[6] S. Bobba, A. Chakraborty, O. Thomas, *et al.*, "CELONCEL: Effective Design Technique for 3-D Monolithic Integration Targeting High Performance Integrated Circuits," in *Proc. Asia and South Pacific Design Automation Conf.*, 2011.

[7] Y.-J. Lee, D. Limbrick, and S. K. Lim, "Power Benefit Study for Ultra-High Density Transistor-level Monolithic 3D ICs," in *Proc. Design Automation Conf.*, 2013.

[8] S. K. Samal, D. Nayak, M. Ichihashi, *et al.*, "Tier Partitioning Strategy to Mitigate BEOL Degradation and Cost Issues in Monolithic 3D ICs," in *Proc. Int. Conf. on Computer-Aided Design*, 2016.

[9] K. Acharya, K. Chang, B. W. Ku, *et al.*, "Monolithic 3D IC Design: Power, Performance, and Area Impact at 7nm," in *Proc. Int. Symp. on Quality Electronic Design*, 2016.

[10] M. G. Bardon, P. Raghavan, G. Eneman, *et al.*, "Group IV Channels for 7nm FinFETs: Performance for SoCs Power and Speed Metrics," in *Proc. Symp. on VLSI Technology*, 2014.

[11] G. Lopez, R. Murali, R. Sarvari, *et al.*, "The Impact of Size Effects and Copper Interconnect Process Variations on the Maximum Critical Path Delay of Single and Multi-Core Microprocessors," in *Proc. Int. Interconnect Technology Conf.*, 2007.

[12] O. v. d. Straten, X. Zhang, K. Motoyama, *et al.*, "ALD and PVD Tantalum Nitride Barrier Resistivity and Their Significance in Via Resistance Trends," *ECS Trans.*, vol. 64, no. 9, pp. 117–122, 2014.

[13] F. Liu, B. Fletcher, E. A. Joseph, *et al.*, "Subtractive W Contact and Local Interconnect Co-Integration (CLIC)," in *Proc. Int. Interconnect Technology Conf.*, 2013.

[14] S. Y. Wu, C. Y. Lin, M. C. Chiang, *et al.*, "A 16nm FinFET CMOS Technology for Mobile SoC and Computing Applications," in *Proc. Int. Electron Devices Meeting*, 2013.

[15] S. Sinha, G. Yeric, V. Chandra, *et al.*, "Exploring Sub-20nm FinFET Design with Predictive Technology Models," in *Proc. Design Automation Conf.*, 2012.

[16] S. Sinha, B. Cline, G. Yeric, *et al.*, "Design Benchmarking to 7nm with FinFET Predictive Technology Models," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2012.

[17] H. Esmaeilzadeh, E. Blem, R. St. Amant, *et al.*, "Dark Silicon and the End of Multicore Scaling," in *Proc. Int. Symp. on Computer Architecture*, 2011.

[18] W. Arden, M. Brillouët, P. Cogez, *et al.*, ""More-than-Moore" White Paper," p. 31,

[19] O. Billoint, H. Sarhan, I. Rayane, *et al.*, "A Comprehensive Study of Monolithic 3D Cell on Cell Design Using Commercial 2D Tool," in *Proc. Design, Automation and Test in Europe*, 2015.

[20] S. H. Yang, J. Y. Sheu, M. K. Ieong, *et al.*, "28nm Metal-Gate High-K CMOS SoC Technology for High-Performance Mobile Applications," in *Proc. Custom Integrated Circuits Conf.*, 2011.

[21] T. Song, W. Rim, J. Jung, *et al.*, "A 14 nm FinFET 128 Mb SRAM With Vrm MIN Enhancement Techniques for Low-Power Applications," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 158–169, 2015.

[22] K. I. Seo, B. Haran, D. Gupta, *et al.*, "A 10nm Platform Technology for Low Power and High Performance Application Featuring FINFET Devices with Multi Workfunction Gate Stack on Bulk and SOI," in *Proc. Symp. on VLSI Technology*, 2014.

[23] P. Batude, C. Fenouillet-Beranger, L. Pasini, *et al.*, "3DVLSI with CoolCube Process: An Alternative Path to Scaling," in *Proc. Symp. on VLSI Technology*, 2015.

[24] P. Larsson, "Resonance and Damping in CMOS Circuits with On-Chip Decoupling Capacitance," *IEEE Trans. on Circuits and Systems*, vol. 45, no. 8, pp. 849–858, 1998.

[25] S. Pant and E. Chiprout, "Power Grid Physics and Implications for CAD," in *Proc. Design Automation Conf.*, 2006.

[26] N. H. Khan, S. M. Alam, and S. Hassoun, "Power Delivery Design for 3-D ICs Using Different Through-Silicon Via (TSV) Technologies," *IEEE Trans. on VLSI Systems*, vol. 19, no. 4, pp. 647–658, 2011.

[27] S. K. Samal, K. Samadi, P. Kamal, *et al.*, "Full Chip Impact Study of Power Delivery Network Designs in Monolithic 3D ICs," in *Proc. Int. Conf. on Computer-Aided Design*, 2014.

[28] S. Das, P. Whatmough, and D. Bull, "Modeling and Characterization of the System-Level Power Delivery Network for a Dual-Core ARM Cortex-A57 Cluster in 28nm CMOS," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2015.

[29] B. Bozorgzadeh and A. Afzali-Kusha, "Decoupling Capacitor Optimization for Nanotechnology Designs," in *Proc. Int. Conf. on Microelectronics*, 2008.

[30] L. Deng, G. Hinton, and B. Kingsbury, "New Types of Deep Neural Network Learning for Speech Recognition and Related Applications: An Overview," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2013.

[31] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," *arXiv:1303.5778 [cs]*, 2013.

[32] A. Conneau, D. Kiela, H. Schwenk, *et al.*, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," *arXiv:1705.02364 [cs]*, 2017.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proc. Int. Conf. on Neural Information Processing Systems*, 2012.

[34] K. He, X. Zhang, S. Ren, *et al.*, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, 2015.

[35] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664–676, 2017.

[36] W. Xiong, J. Droppo, X. Huang, *et al.*, "The Microsoft 2016 Conversational Speech Recognition System," *arXiv:1609.03528 [cs]*, 2016.

[37] W. Liao, L. He, and K. M. Lepak, "Temperature and Supply Voltage Aware Performance and Power Modeling at Microarchitecture Level," *IEEE Transactions on*

Computer-Aided Design of Integrated Circuits and Systems, vol. 24, no. 7, pp. 1042–1053, 2005.

[38] V. Sze, Y.-H. Chen, J. Emer, *et al.*, "Hardware for Machine Learning: Challenges and Opportunities," *arXiv:1612.07625 [cs]*, 2017.

[39] T. He, Y. Fan, Y. Qian, *et al.*, "Reshaping Deep Neural Network for Fast Decoding by Node-Pruning," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2014.

[40] S. Han, H. Mao, and W. J. Dally, "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding," *arXiv:1510.00149 [cs]*, 2015.

[41] D. Kadetotad, S. Arunachalam, C. Chakrabarti, *et al.*, "Efficient Memory Compression in Deep Neural Networks Using Coarse-Grain Sparsification for Speech Applications," in *Proc. Int. Conf. on Computer-Aided Design*, 2016.

[42] Y. Cheng, D. Wang, P. Zhou, *et al.*, "A Survey of Model Compression and Acceleration for Deep Neural Networks," *arXiv:1710.09282 [cs]*, 2017.

[43] M. Courbariaux, Y. Bengio, and J.-P. David, "BinaryConnect: Training Deep Neural Networks with Binary Weights During Propagations," *arXiv:1511.00363 [cs]*, 2015.

[44] M. Courbariaux, I. Hubara, D. Soudry, *et al.*, "Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1," *arXiv:1602.02830 [cs]*, 2016.

[45] D. Su, X. Wu, and L. Xu, "GMM-HMM Acoustic Model Training by a Two Level Procedure with Gaussian Components Determined by Automatic Model Selection," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2010.

[46] J. S. Garofolo, L. F. Lamel, W. M. Fisher, *et al.*, "DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 1993.

[47] D. Povey, A. Ghoshal, G. Boulianne, *et al.*, "The Kaldi Speech Recognition Toolkit," *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[48] W. A. Gardner, "Learning Characteristics of Stochastic-Gradient-Descent Algorithms: A General Study, Analysis, and Critique," *Signal Processing*, vol. 6, no. 2, pp. 113–133, 1984.

[49] S. Han, J. Kang, H. Mao, *et al.*, "ESE: Efficient Speech Recognition Engine with Sparse LSTM on FPGA," *arXiv:1612.00694 [cs]*, 2016.

[50] Y. Cheng, F. X. Yu, R. S. Feris, *et al.*, "An Exploration of Parameter Redundancy in Deep Networks with Circulant Projections," *arXiv:1502.03436 [cs]*, 2015.

[51] S. Liao, Z. Li, X. Lin, *et al.*, "Energy-Efficient, High-Performance, Highly-Compressed Deep Neural Network Design Using Block-Circulant Matrices," in *Proc. Int. Conf. on Computer-Aided Design*, 2017.

[52] S. Gray, A. Radford, and D. P. Kingma, "GPU Kernels for Block-Sparse Weights," OpenAI, Tech. Rep., 2017.

PUBLICATIONS

This dissertation is based on and/or related to the works presented in the following publications in print:

[1] K. Chang, K. Acharya, S. Sinha, *et al.*, "Power Benefit Study of Monolithic 3D IC at the 7nm Technology Node," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2015.

[2] K. Chang, S. Sinha, B. Cline, *et al.*, "Match-Making for Monolithic 3D IC: Finding the Right Technology Node," in *Proc. Design Automation Conf.*, 2016.

[3] K. Chang, S. Sinha, B. Cline, *et al.*, "Cascade2D: A Design-Aware Partitioning Approach to Monolithic 3D IC with 2D Commercial Tools," in *Proc. Int. Conf. on Computer-Aided Design*, 2016.

[4] K. Chang, K. Acharya, S. Sinha, *et al.*, "Impact and Design Guideline of Monolithic 3-D IC at the 7-nm Technology Node," *IEEE Trans. on VLSI Systems*, vol. 25, no. 7, pp. 2118–2129, 2017.

[5] K. Chang, S. Das, S. Sinha, *et al.*, "Frequency and Time Domain Analysis of Power Delivery Network for Monolithic 3D ICs," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2017.

[6] K. Chang, D. Kadetotad, Y. Cao, *et al.*, "Monolithic 3D IC Designs for Low-Power Deep Neural Networks Targeting Speech Recognition," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2017.

[7] K. Chang, S. Pentapati, D. E. Shim, *et al.*, "Road to High-Performance 3D ICs: Performance Optimization Methodologies for Monolithic 3D ICs," in *Proc. Int. Symp. on Low Power Electronics and Design*, 2018.

[8] K. Chang, D. Kadetotad, Y. Cao, *et al.*, "Power, Performance, and Area Benefit of Monolithic 3D ICs for On-Chip Deep Neural Networks Targeting Speech Recognition," *ACM Journal on Emerging Technologies in Computer Systems*, vol. 14, no. 4, 42:1–42:19, 2018.

[9] K. Chang, S. Das, S. Sinha, *et al.*, "System-Level Power Delivery Network Analysis and Optimization for Monolithic 3-D ICs," *IEEE Trans. on VLSI Systems*, pp. 1–11, 2019.

In addition, the author has completed works unrelated to this dissertation presented in the following publications in print:

[1] K. Acharya, K. Chang, B. W. Ku, *et al.*, "Monolithic 3D IC Design: Power, Performance, and Area Impact at 7nm," in *Proc. Int. Symp. on Quality Electronic Design*, 2016.

[2] K. Chang, A. Koneru, K. Chakrabarty, *et al.*, "Design Automation and Testing of Monolithic 3D ICs: Opportunities, Challenges, and Solutions," in *Proc. Int. Conf. on Computer-Aided Design*, 2017.

[3] K. Chang, B. W. Ku, S. Sinha, *et al.*, "Full-Chip Monolithic 3D IC Design and Power Performance Analysis with ASAP7 Library: (Invited Paper)," in *Proc. Int. Conf. on Computer-Aided Design*, 2017.

[4] B. W. Ku, K. Chang, and S. K. Lim, "Compact-2D: A Physical Design Methodology to Build Commercial-Quality Face-to-Face-Bonded 3D ICs," in *Proc. Int. Symp. on Physical Design*, 2018.

VITA

Kyungwook Chang was born in Compiègne, France, in 1985. He received his B.S. degree in Electrical and Computer Engineering in 2007 from Seoul National University, S. Korea, where he also received M.S. degree in Electrical Engineering and Computer Science in 2010.

He has been working as a graduate research assistant in GTCAD Laboratory housed in the School of Electrical and Computer Engineering at Georgia Institute of Technology from 2014. His primary research interests are CAD and physical design solutions for 3D IC design challenges, design and technology co-optimization. His other research interests include high-performance and low-power designs, and parallel computing and memory architecture.